The use of the Qualitative Trajectory Calculus in sports analytics

Jasper Beernaerts

The research reported in this dissertation was conducted at the CartoGIS Research Unit, Department of Geography, Faculty of Sciences, Ghent University.

Faculty of Sciences
Department of Geography

The use of the Qualitative Trajectory Calculus in sports analytics

Thesis submitted in accordance with the requirements for the degree of Doctor of Sciences: Geography

Het gebruik van de Kwalitatieve Traject Calculus in sport analytics

Proefschrift aangeboden tot het behalen van de graad van Doctor in de Wetenschappen: Geografie

Beernaerts Jasper

Funded by FWO-Vlaanderen grant no. 11ZN118N

Promotors:

Prof. dr. Nico Van de Weghe
Department of Geography, Ghent University

Prof. dr. Bernard De Baets
Department of Data Analysis and Mathematical Modelling, Ghent University

Prof. dr. Matthieu Lenoir
Department of Movement and Sports Sciences, Ghent University

Members of the examination committee:

Prof. dr. Veerle Van Eetvelde (chair)
Department of Geography, Ghent University

Prof dr. Christophe Claramunt
French Naval Academy, Brest Naval

Prof. dr. Frederik Deconinck
Department of Movement and Sports Sciences, Ghent University

Prof. dr. Guy de Tré
Department of Telecommunications and Information Processing, Ghent University

Prof. dr. Nikolaos Mavridis
Interactive Robots and Media Lab, United Arab Emirates University

Prof. dr. Roel Vaeyens
Department of Movement and Sports Sciences, Ghent University

Prof. dr. Erik Witvrouw
Department of Rehabilitation Sciences, Ghent University

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I still remember my first days as a Bachelor student at the Department of Geography as it was the day of yesterday. While I had chosen the topic of Geography solely based on my general interests, I had no clear view of my professional ambitions at that time. Adapting myself to the study rhythm at the University - not a straightforward task for someone who was used to rehearse all his classes the same day - I had the chance of exploring and enjoying the rich social activities that came along with studying at this small but cosy department. After three years of studying Geography at Ghent University, I entered my Masters still wondering what I would become later on in my professional life. Browsing through the options of specialisation in the Masters, I quickly decided to specialize myself in 'GIScience', and thereby acquired a taste for big data analysis and spatial analytics. After all, I had always been fascinated by moving objects, simulating traffic streams on a small scale from since I was a little boy. Back in those days, I often built entire cities in my parents' house's living room, including miniature railways, cars, ships and even airfields taking passengers of my town to their far away destinations. Later, when my mom finally convinced me that she needed the living room for other purposes, I switched to simulation games on the computer. Many years later, however, I discovered the pleasure and satisfaction of conducting research in a domain of your own interest, when writing my master thesis on Bluetooth-tracking of pedestrians in the city of Ghent.

Besides my interest for moving objects and their interactions, I have been a sports enthusiast for many years, playing badminton, basketball, tennis and swimming at professional levels. Furthermore, I have been following various football competitions very closely since my childhood, and I often went out playing football with my friends at the local playground. I remember the countless cosy evenings of watching and commenting on matches of the 'Champions League' in front of the fireplace with my father. Immense was my enthusiasm and gratitude when Prof. dr. Nico Van de Weghe offered me the opportunity to work as a scientific researcher on a project titled "*Automated analysis of sports team movement tactics: development of a demonstrator*

and initiation of a license track" in the summer of 2014. During these months as scientific researcher at Ghent University, I got a taste of the life as a researcher, combining my interest in sports with my passion for spatial analytics. In fact, this period was so pleasant and inspiring that I decided to apply for a PhD grant at the FWO Research Foundation Flanders later that year, to conduct more fundamental research on this topic. Knowing chances to obtain a grant were limited and competition was fierce, I once more would like to express my sincere gratitude to the FWO for accepting my proposal and allowing me to conduct research on the use of Geographical models for Sports Analytics. After receiving the grant from the FWO, I started working on my PhD in October 2015.

Indeed, the past four years have been extremely interesting and enriched me both on a personal as well as on a professional level. For this, I would like to thank my promotors, Prof. dr. Nico Van de Weghe, Prof. dr. Bernard De Baets and Prof. dr. Matthieu Lenoir from the bottom of my heart. They have been truly great mentors along the road, always providing insights and critical feedback, each from their own field of specialization. I really appreciate that Prof. Van de Weghe gave me the opportunity to guide and support 8 master students with conducting their Master thesis research during my stay at Ghent University. Moreover, I would like to show my sincere gratitude to dr. Kristof De Mey, who saved time nor effort to help me get in touch with sports teams, companies and sports professionals. I can honestly say that without him it would not have been possible to conduct my research on this level. Furthermore, I would like to thank BC Oostende, Deltatre and the UEFA for supporting my research by providing sports data. My gratitude goes out to all the co-authors of the papers I was able to write or contribute to. Last but not least, I would like to thank my colleagues at the Department of Geography, who transformed the workplace into a welcoming and relaxing haven. As I saw many colleagues come and go, I would like to thank all of them for making the past years unforgettable both on and off the job. I can happily say that many of them have become good friends. My gratitude goes out to Helga who spent hours helping me with the inevitable administrative tasks and gave wonderful advice and comfort during the past years.

On a personal level, I want to thank all my close friends, for providing the necessary spare time off the job, as well as -for my more sports minded friends- the hour-long debates about, as they call it, ‘my inventions’. Despite being able to thank all of my friends personally by name, I would like to thank Yasin, Andries, Axel, Frederick, Sarah and Berkay. Indeed, you are truly the best friends and supporters I could have ever wished for. In particular, I would like to show my gratitude to Axel, who was my study companion throughout my university career. Axel, you showed me how exciting acquiring knowledge and critically discussing them with others can be. To my parents, Patrick and Ingrid and my sister, Sien, I would like to express my sincere gratitude. Growing up in such a warm family, receiving your lifelong and unconditional support, has enabled me to become the man I am today. At this moment, I would like to thank the wonderful woman I met during the first year of my career at Ghent University. Duygu, your love, relentless support, encouragements and understanding have given me the force to finish this dissertation. Without you I would not have found the patience to fight through the hard moments that came along with conducting this research.

I would like to conclude by looking back at two pieces of advice I was given a couple of years ago, ‘*Work to live, don’t live to work*’ and ‘*To achieve happiness, follow your passions*’. After writing this dissertation I can indeed say that, once your passion becomes your work, little effort is needed to happily combine work and private life. For this I would like to thank everybody who helped me along the way.

Jasper Beernaerts

Gent, June 2019

DECLARATIONS

Parts of this work are based on papers previously published in:

Beernaerts, J., De Baets, B., Lenoir, M., De Mey, K., Van de Weghe, N. (2018). Analysing team formations in football with the static qualitative trajectory calculus. In P. Pezarat-Correia, J.P. Vilas-Boas, O. Rivera & J. Cabri (Eds), Proceedings of the 6th International Congress on Sport Sciences Research and Technology Support (pp.15-22). Seville, Spain: Scitepress.

Beernaerts, J., Derie, R., Nguyen, B., Vansteenkiste, P, De Baets, B., Deconinck, F.J.A., Lenoir, M., De Clercq, D., Van de Weghe, N. (2018). Assessing the potential of the qualitative trajectory calculus to detect gait pathologies: a case study of children with development coordination disorder. *Computer Methods in Biomechanics and Biomedical Engineering*, 1-7.

Beernaerts, J., De Bever, E., Lenoir, M., De Baetts, B., Van de Weghe, N. (2019). A method based on the Levenshtein distance metric for the comparison of multiple movement patterns described by matrix sequences of different length. *Expert Systems with Applications,* 115, 373-385.

Zhang, P.D., Beernaerts, J., Zhang, L., Van de Weghe, N. (2016). Visual exploration of match performance based on football movement data using the continuous triangular model. *Applied Geograpphy*, 76, 1-13.

Zhang, P.D., Beernaerts, J., Van de Weghe, N. (2018). A hybrid approach combining the multi-temporal scale spatio-temporal network with the continuous triangular model for exploring dynamic reactions in movement data: a case study of football. *ISPRS International Journal of Geo-Information*, 7(1), 1-18.

I would like to thank all the co-authors for their contributions

ABSTRACT

The domain of sports analytics has seen a massive growth in the last decennium. In the subdomain of sports performance analysis, various methods were introduced, going from psychological and sociological assessments of sports teams to statistics of individual athletes such as respiration rates, heartbeat ratios or information on the nutrient intake. Despite the huge variety of sports with very different characteristics, most sports have a common component: (spatial) movement through walking or running. This dissertation presents the research on the use of the Qualitative Trajectory Calculus (QTC) for sports analytics. QTC is a method that originated from the field of Geography that describes the relative movement between Moving Point Objects (MPOs) by means of qualitative symbols. The goal of the research is the development of a methodology for (automated) sports pattern recognition through the analysis of complex movement interactions between different players based on the QTC. Such a method could aid trainers, coaches and sports performance analysts with analysing the performed movement patterns in (team) sports. As movement patterns in sports can be analysed using both a fine temporal resolution as well as a more coarse temporal resolution, the distinction is made between a method for dynamic sports analytics and a method for static sports analytics. The method for dynamic sports analytics makes use of traditional QTC-variants and aims at analysing team behaviour through the description and comparison of movement patterns of players with a fine temporal resolution. The static sports analytics method, on the other hand, analyses the spatial configuration of the (averaged) positions of players and aims at describing team tactics. The static method makes use of a novel QTC-variant. Both methods are described rigorously in this dissertation and their characteristics are described by a series of three-cushion billiard examples. Three cushion billiards is a variant of the more popular carom billiards and was chosen as an example because of its simplicity. In three-cushion billiards, the number of objects is limited and the movements of the balls are only defined by an initial shot and the laws of nature. The absence of psychological or human factors influencing the trajectories of the balls after the shot has been taken reduces the complexity and allows to evaluate the use of the proposed methods for dynamic and

static sports analytics before switching to more complex sports situations. To conclude the methodology section, the implementation of both methods by means of two Python programs is described. Besides the application of the methods for analysing team sports, their use for analysing more detailed movements of individual athletes was investigated. Three use cases with each a clear set-up, a representative dataset and a rigorous (successful) validation are presented, to evaluate the use of the dynamic and static sports analytics methods for real sports cases. The first use case presents an application of the dynamic sports analytics method for the recognition of basic spatial movement patterns in football. It shows that the proposed method allows a user to select a movement pattern as a reference and to search for similar movement patterns in one or multiple football matches. The second use case demonstrates that the proposed method for dynamic sports analytics can be used to analyse the more detailed movements of parts of the human body. More specific, it allows for the detection of gait pathologies, such as the Development Coordination Disorder (DCD), by describing and comparing the walking patterns of multiple individuals with QTC. The third use case evaluates the use of the static sports analytics method for analysing team tactics in football. It shows that the method can be used to describe and analyse the team formations of a whole team or groups of players in football. After the use cases, a thorough discussion and conclusion of the results are presented.

1.

INTRODUCTION

The sports analytics market has seen a tremendous growth in recent years, with a market valuation growing from $125 million in 2014 to a prospected $4.7 billion in 2021 (Reportstack, 2015). While the market is diverse and provides analytics for a huge variety of sports, this growth is being reflected into an emerging interest of the academic world for sports sciences and sports analytics (Coleman, 2012). Moreover, media reports and movies such as *Moneyball* (De Luca *et al.*, 2011), starring Brad Pitt as a sports analytics early adopter and later on sports statistics guru in baseball, raised the awareness of the big public for the impact of sports analytics on sports at all (even amateur) levels and, thus, life in general. Indeed, the movie *Moneyball* is inspired on the real-life story of Billy Bean, who is seen as a pioneer for using statistical analyses for player recruitment in baseball. Sports analytics thus became increasingly known in recent years, through various sports. In 2015, for example, popular media picked up the story about a small Danish football[1] club, Midtjylland, who had miraculously qualified itself for the UEFA Champions League main tournament despite a minimal budget, through the extensive use of novel sports analytics (Ingle, 2015). Sports analytics as such rapidly became a hot topic and buzzword, with researchers from multiple disciplines doing efforts on developing new methods and techniques. These disperse efforts, however, resulted in the lack of a clear and generally accepted definition of sports analytics (Wright, 2009). Based on this, I am convinced that at this time, few of the readers of this manuscript have a good overview of the sports analytics market and its scope. As such, this dissertation starts with an introduction to the sports analytics market, to better situate

[1] Please note that this dissertation is written in British English, as such the term ‘football’ is used rather than the American term ‘soccer’.

my research in this context. Since this dissertation includes novel research on various sports such as billiards, golf, running, basketball and football, the literature overview will consequently include examples of different sports.

1.1 Introducing sports analytics

In this section, the domain of sports analytics is introduced. After defining the boundaries of the domain, a general overview of the state of art in sports analytics is provided, both from an academical as well as from an commercial viewpoint.

1.1.1 Defining sports analytics

Sports analytics methods try to go beyond the 'gut' feeling of a coach or player for optimizing sports performance, by providing objective measurements of a player or teams' sports performance. While multiple definitions of sports analytics exist, an often mentioned one is the definition from Almar (2013): "*The management of structured historical data, the application of predictive analytical models that utilize that data, and the use of informational systems to inform decisions makers and enable them to help their organizations in gaining a competitive advantage on the field of play*". Countless variations on this definition exist, some of them focussing on data mining and knowledge discovery by visualisation of sports data (Snijders *et al.*, 2012; Siegel, 2013). Others, on the other hand, put more emphasis on the predictive character of sports analytics (Shmueli, 2010; Waller & Fawcett 2013). Swartz (2018), however, accurately states that, because of the diverse background of the practitioners, every attempt to define sports analytics is bound to cause disagreement. First of all, there is no consensus on what a sport is.

1.1.2 What is sports?

Efforts were made by Levinson & Christensen (1996) to list up and describe all sports in an encyclopaedia, by defining sports based on two principles. First, according to them, sports should include some form of competition, where the winner is appointed based on fixed rules. Second, the primary goal of sports consists of winning, and the chances of winning should be based on the relative physical ability of the competitors, even though strategy and chance could also play a (minor) role. Following this rather inclusive definition, Levinson and Christensen distinguished more than 300 sports. The above-mentioned definition of sports implies that new sports can be invented and could be added to this list. Up to 1990, most of the sports had been invented in the Western world, but from then on, a rise of inventions of new sports can be seen all over the world (Whannel, 2005). In a similar attempt to list up and describe sports less than a decade later, Liponski (2003) distinguished more than 8000 different sports. At the moment of writing, it is clear that new sports will keep on being invented in the future. Moreover, rules of established sports, as described by Levinson & Christensen (1996) and Liponski (2003), can be changed through time, as happened with football with the introduction of the penalty spot in 1891 (Squires, 2016) or, more recently, changes to the scoring system in badminton in 2005 (He-zhou, 2006). Another example is the decision to change the diameter of the official table tennis balls from 38 mm to 40 mm, to make the rallies slower and the sports more attractive for television broadcasting (Hughes, 2007).

1.1.3 Analytics in sports

Analytics in general can be defined as "*The process of developing actionable insights through problem definition and the application of statistical models and analysis against existing and/or simulated future data*" (Cooper, 2012). Applying this definition of analytics to the domain of sports, sports analytics could include multiple subdomains, e.g. sports fan analytics or fan engagement analytics (Yoshida *et al.,* 2014), sports marketing analytics (Shwarz & Hunter, 2012) and sports performance analytics (O'Donoghue, 2015). With respect to these, it must be noted that this dissertation presents a method that is a part of the latter subdomain, *i.e.* sports performance analytics. As such, the following sections will not discuss the full scope of sports analytics as defined above.

Besides the lack of good definitions for sports and to a lesser extent, analytics, there is the problem that almost an endless number of factors influence sports performance. Assume, for example, that there is a sports analytics method aimed at measuring and modelling the acceleration of a sprinter. Surely muscle strength will be an influencing factor for his/her performance, yet psychological issues might, to a certain extent, also influence a runner’s acceleration. At that point, the question arises whether psychological analysis methods should be considered to be a sports analytics method or not. Due to such dilemmas, sports analytics remains an ill-defined domain.

1.1.4 Academic and commercial research on sports analytics

Sports analytics methods started appearing more than 50 years ago, but the highest numbers of introduced methods can be situated in more recent years. Numbers of journals and papers discussing the topic of sports analytics substantially rose from 1990 on, supported by technological developments and rising computational power (Coleman, 2012). The lack of a good definition for sports analytics (Section 1.1.1), however, makes it somewhat difficult to pinpoint its first occurrence in history. Often mentioned as the first occurrence are the coded notes in American football and basketball in 1960 (Franks & Hughes, 2004). The notes were used to objectively register sports performance and events during official matches, and later on enabled researchers to conduct analyses on the gathered data. While those analyses were very rudimentary (e.g. counting the number of passes or scored points per player), they are considered as the first organised attempt of analysing sports data. The term ‘organised’ here means that the method is (or was) used by multiple analysts and for multiple teams. Before that in Europe, a region where football traditionally was more popular, Charles Reep (Morgulev *et al.*, 2017) started counting the number of passes that led to a goal in football, and their position of occurrence on the field. While his method was introduced in 1950, it is not commonly considered as the first generally accepted sports analytics method, as it was only adopted by few other analysts at later times.

Research in sports analytics at those times, however, was constrained by effort-intensive methods for sports data collection and limited computational power (the lack of cheap and powerful computers) for analysing the gathered data (Morgulev *et al.*, 2017). In that perspective, it is not surprising that analytics were often first applied to so-called 'fantasy sports'. Fantasy sports are (online) computer games that have the aim of mimicking real sports, e.g. the characteristics of the players, the rules of the game. They are increasingly popular with sports fans and are often combined with betting. Since fantasy sports are computer simulations, data can be logged and analysed more easily than the real sports they try to mimic (Bonomo *et al.*, 2014). Since in recent times the problem of data acquisition is quickly disappearing, a wide-spread growth of academic interest in sports analytics can be observed, with a rising number of universities offering courses and even bachelor programmes and majors in sports analytics.

Besides the emerging interest of the academic world in sports analytics (Coleman, 2012), a growing number of commercial companies are conducting research and providing services in the domain of sports analytics. Examples of famous companies are, amongst many others, STATS (e.g. American football, basketball, football, hockey, rugby), OPTA (notational analysis in football, rugby and others), Gracenote (football), Catapult Sports (wearables for athletes), Deltatre (e.g. football), Hawk-Eye (e.g. badminton, football, snooker, tennis, volleyball) and Longomatch (notational analysis in football for e.g. the Belgian national football team). Furthermore, these companies deliver data to media companies, which try to use fan engagement as a tool for gaining more visibility and ultimately, more readers or viewers. In that respect, it is vital to have the latest insights, statistics and details about the fans favourite competition, team and players. Closely linked to the sports media and fantasy sports is the betting on (real) sports matches.

To facilitate communication and the exchange of ideas between both ends of the spectrum, various sports analytics conferences were established in the past decades, some of them positioned at the commercial end (e.g. the annual MIT Sloan Sports Analytics Conference, and the Sports Analytics World Series that are held multiple times a year in different continents) and others more with a focus on academic research (e.g. the annual International Conference on Sport Science and Technology (icSPORTS), and the annual International Conference on Physical Education and Sports Science (ICPESS)). Nevertheless, all of these conferences are characterised by far-stretching interdisciplinarity and a rising popularity. As a consequence, universities are trying to establish a direct communication with the sports analytics industry, to ensure a good link and trade-off between sports research and its industrial value. An example of this is the Victoris Consortium at Ghent University, which facilitated the research presented in this dissertation. The question is, however, to what extent commercial companies or professional teams are enthusiastic about sharing their sports analytics practice with the (scientific) world. Following the unexpected success of the earlier mentioned football team Midtjylland, the club owner publicly stated that their success was based on the application of sports analytics. A key component of their sports analytics approach was the analysis of set pieces (e.g. corners and free kicks). Their efforts for optimizing the team performance on set pieces resulted in 25 goals from set pieces in the season 2014-2015, which was more than double the number of goals from set pieces of the second-best performing team in the Danish first league (11). As other teams witnessed this success of sports analytics practices in football, they quickly adopted their own, often very similar, sports analytics principles. In this way, Midtjylland lost a big part of this competitive advantage, as the other teams of the Danish competition (and later of the whole world) copied their practice.

1.1.5 Differences between sports

Sports analytics in basketball has evolved during more than half a century to a point where the NBA, the world's most popular basketball competition, has embraced analytics and embedded them into the decision-making process of the teams and the league. Together with baseball, another popular sports in Northern America, it has set the standards for professionalization and integration of sports analytics in official sporting competitions. In the NBA, for instance, a regulation was made (Skinner & Guy, 2015) so that since the season of 2013-2014, all of the stadiums should be equipped with a system that automatically logs and processes the players coordinates during the game. As these data are distributed to all the teams in the competition, this created a platform for sports analytics to develop further. Federations of other sports such as football and hockey, on the other hand, have traditionally been less keen on introducing technological evolutions such as wearables, goal line technology, etc. In the last decade, however, it can be observed that even those federations are adopting policies for applying sports analytics in official matches (D'Orazio and Leo, 2010; Link, 2018). In 2018, for example, the Belgian Jupiler Pro league (the highest professional football league in Belgium) contracted the commercial company STATS to log data from all of the matches and to create a platform in which competition game data from all professional clubs are being shared. This shows that the difference in levels of adoption of sports analytics in the different sports is closely linked to the issue of data availability. The technology that was introduced in the NBA can indeed be easily used for football to track the players positions during the match. In other sports however, such as judo, taikwando (Marcon *et al.*, 2010) or golf, it would be much more interesting to track specific body parts rather than obtaining one set of coordinates for each individual player (Fung *et al.*, 2014). This, however, is not a straightforward task due to player occlusions and rapid movement of the players. Yet, in other sports, such as swimming, wearables seem to be the only solution to acquire accurate (positional) data. To efficiently acquire such data, it is important for federations to regulate the use of tracking technologies and to promote the use of standardized data formats.

Once the data is acquired, questions regarding the data ownership arise. Since in team sports individual players are being contracted and paid by their teams, it seems logical that their teams should become the owners of the data. After all, they are the ones who paid for the data collection. However, once a player is transferred from one team to another, some questions arise regarding the ownership of the data. Can the transferred player take his/her data to his or her new team? Should the former team delete the old data? Indeed, many questions arise regarding data ownership in sports. Furthermore, data is often sensitive and consists of privacy-sensitive information regarding the health of the player. The debate becomes even more complex when considering the fact that teams often buy licences from companies that gather the sports data. This means that they are provided with data upon request, implying they actually never become owner of the data. For individual sportsmen or women, such as athletes or recreational sportsmen or women, data ownership seems much more straightforward. Few regulations regarding data ownership exist in the world of sports at the present time. As data gathering in sports became easy, only very recently, we expect policies to be adapting in the future to this new reality. Governmental efforts on data regulation, such as the General Data Protection Regulation (GDPR), which was implemented by the EU government in 2018, might speed up regulations regarding sports data (Keane, 2018). It must be said that these issues can make it very hard for researchers to obtain good data, as was the case for the research presented in this dissertation.

1.1.6 Limitations of sports analytics and the impact on game experience

We would like to end this chapter with an important thought, by using a saying derived from the theories of the economist Goodhart (1981), "When a measure becomes a target, it ceases to be a good measure". It is important for people who are dealing with sports analytics to keep in mind that everything starts with the beauty of the game and statistics should be used as a way to obtain more insight into the game rather than to be a goal of its own. It is indeed tempting to apply statistics and design analytics to every aspect of the game to try and optimize it. However, one should not forget the core and logic of the game. In football, for example, the number of penalties awarded to a team might be a good indicator of the team's expected scored goals and thus performance. While there is indeed a causal relation in this respect, it would be unwise for a coach to focus training on provoking penalty fouls and, for the strikers, to shoot the perfect penalty. It is not difficult to see that it would be better to train the team's general attacking skills and train the strikers to shoot to the goal from different positions and angles, and not only from the penalty spot. Yet, unwise use of sports statistics to support major decisions in sports are prevalent. The transfer policy of Liverpool FC, for example, became heavily reliant on sports statistics in 2011, as the club was bought by the American Fenway Sports Group the year before. The company previously owned multiple baseball teams and was known to make extensive use of sports statistics for transfers of baseball players. As sports statistics showed that Liverpool FC owned the best winger of the English Premiere League at that time, they decided to buy Andy Caroll, the football player with the most header goals of that season. Despite the theoretical perfect combination of the best winger and best header, the transfer proved to be very unsuccessful in reality. Scoring very few goals, Andy Caroll was sold again little over a year later, as statistics showed that, on average, 18 crosses were needed to score one header goal in the Premiere League. Due to this expensive mistake, Fenway Sports Group quickly learnt that sports statistics that are useful for baseball are not necessarily applicable to football, as the latter is characterized by a higher unpredictability.

Besides the limited applicability of sports analytics, there are some moral questions regarding sports analytics. More and more people, for example, believe that extensive use of analytics in sports takes away the joy and the beauty of the game (Lengel, 2018; McCaskill, 2018; Timms, 2018). After all, isn't sports meant to be a leisure activity, something that should be accessible to everyone, to forget about the daily life struggles? Although this is in a way correct, one could clearly state that professional football players are under an immense pressure and stress for the games they play, whether or not the game would be analysed. One way to see it is that the introduction of sports analytics methods are inevitable as the popularity of a sports rises. Because of their popularity, professional (football) teams have become companies on their own, and some of the world's most popular teams (e.g. Juventus, Manchester United and AS Roma) are now listed on the stock market. As sportive results affect the stock price of the share or the return of investment of private investors (through constructions including holdings and shareholders), teams will do everything they can to make sure they win games, and therefore look at sports analytics to enhance the team's performance (Benkraiem *et al.*, 2009). The stock market value of the Italian football team Juventus, for example, soared with a massive 95% in the days following the purchase of the Portuguese star player Cristiano Ronaldo in the summer of 2018. Similarly, the value of the Juventus shares increased with a staggering 17% after Ronaldo scored a hat-trick thereby qualifying the team for the final 8 of the 2018-2019 UEFA Champions League (Mohamed, 2019). Furthermore, sports analytics can enhance the fan engagement, by providing (live) statistics and insights about players, team or competitions (Stockwell, 2015).

1.2 Research objectives and approach

In this section the objectives of the research are introduced, within the sports analytics framework that was present at the start of the research (early 2015). Because sports analytics is a fast-changing domain, it is important to view the research objectives within the state-of-the-art of sports analytics at that time. Following the discussion of the research objectives, we present the research approach, *i.e.* the chosen path to achieve the chosen research objectives. Besides general conclusions (Chapter 10), Chapter 9 includes a current (early 2019) critical examination of the research objectives with respect to the evolutions that happened in the sports analytics domain in the years during which this research was conducted.

1.2.1 Research objective

In the years preceding 2015, data gathering, analysis and interpretation in the world of sports had seen a staggering growth, a trend that has proven to continue in the following years. At the time of writing of the research objectives, sport teams were spending more resources than ever before trying to optimize their players and teams performance. Capturing player movements had become easier because of improvements in sensor technology which occurred mainly in the beginning of the 21st century. Companies and sports teams, however, were bound in their analysis by the complexity of huge datasets (Reunes, 2014) and the limited amount of computing power. As such, in 2015, the state-of-the-art data analysis of sports performance was limited to the static analysis of (x,y,t)-coordinates of players, thereby not taking into account the complex interactions between the athletes. From the sport sciences perspective, there had been a focus on primarily individual performance characteristics (e.g. distance ran, velocity and sprint efforts, passes given, intake of oxygen, production of sweat) or basic team tactics and strategy (e.g. percentage of ball possession, goal shots, number of corners in football, number of successful passes).

In the world of football, for example, the tool '*Soccerstories*' had been renowned (Perin *et al.*, 2013) for its ability to visualise interesting phases of a match. Bialkowski *et al.* (2014) went beyond this visualisation and analysed the impact of the so-called home advantage of football teams by measuring the average positioning of the players on the field. The calculation of individual and aggregated performance characteristics was described by Duarte *et al.* (2012), where the entropy -of what they called compound positional variables- was calculated. They reasoned that the entropy indicated the complexity of the game, and therefore the fatigue of the players. In volleyball some research had been done on detection and classification of defensive strategies (Seweryniak *et al.*, 2013; Jäger & Schöllhorn, 2007). In basketball, player movement was tracked to discover in what way teams achieved to create a so-called 'open shot' on goal (Lucey *et al.,* 2014). These kinds of analysis resulted in eye-catching heat maps and statistical facts, but failed to capture more interesting patterns that could help with making tactical decisions to increase the team's chance of winning the game (Lucey *et al.*, 2014). Furthermore, research at that time would tend to focus on specific sports and failed to create a generic sports analytics methodology. The motivation for our research therefore came from the idea that a general framework for reasoning about player tracking data had to be established in order to make use of the new rich source of sports information. Considering the huge amount of data and the limited number of established and general accepted methods for sports analytics, it was of uttermost importance to start the research at a relatively simple level, only taking in account the most relevant information first.

In that respect, it is advisable to start analysing complex phenomena by first conducting qualitative research, since this has the capabilities to easily show insights and detect potential problems at early stages of the research (Iwasaki, 1997). It is a fact that, while trying to understand complex phenomena, such as sports interactions, it is not always the goal or feasible to analyse or understand everything there is to know (Van Belleghem *et al.*, 1994; Frank, 1996). Considering this, Cristani *et al.* (2000) argued that qualitative approaches are much more appropriate for handling incomplete and unknown information than their quantitative counterparts. Furthermore, evidence had showed that by describing large quantitative datasets in a qualitative manner, computing requirements could be reduced (Freksa, 1992). As such, the idea was to build a multi-sports qualitative framework requiring only positional sports data. Based on promising earlier studies (Delafontaine *et al.* 2011; Van de Weghe *et al.*, 2005), the belief was that by using the Qualitative Trajectory Calculus (QTC, Van de Weghe, 2004), it would be possible to look into player-player interactions and analyse more complex (movement) relations between players than what was possible with the established methods at that time. Compared with other qualitative calculi, QTC had the advantages of being characterised by its relative simplicity and resemblance with the human psychological processes of relative positioning (for example during sports).

Aiming to contribute to the fast-developing domain of sports analytics, the main research objective of the research presented in this dissertation can be defined as:

The development of a methodology for (automated) sports pattern recognition through the analysis of complex movement interactions between different players based on the Qualitative Trajectory Calculus

Such a methodology could give trainers, coaches, sports performance and game analysts the ability to measure team organization, plan tactics and strategies and evaluate particular team interventions by providing them with (a series of) sports analytics methods based on QTC. Based on this broad main research objective and the motivation of the research towards a qualitative approach, following research questions can be derived:

R1 **Can the Qualitative Trajectory Calculus be used to (automatically) detect frequently occurring movement patterns between a certain number of players? (team behaviour)**

R2 **Can the Qualitative Trajectory Calculus be used to (automatically) detect tactical patterns in sports data? (team tactics)**

R3 **In which aspects of sports analytics could the QTC methodology be useful, besides analysing team behaviour and team tactics?**

The first research question (**R1**) involves sports analytics at a fine temporal resolution, *i.e.* the detection of frequently occurring movement patterns between certain numbers of players. In layman terms this could be explained as the detection of multiple occurrences of similar movements of a group of players during training or during one or multiple matches. Although it is true that for this kind of analysis, data at a fine temporal resolution is needed, the minimum required temporal granularity might be heavily dependent upon the sports that will be subjected to the analysis. For analysing basketball players, for example, a finer temporal resolution would be required than when analysing cyclists, since the first usually move with higher degrees freedom. Consequently, the analysis discussed in **R1** can be referred to as the analysis of **team behaviour**, as it focuses on movement interactions between the players of limited sets of players during a limited amount of time.

The second research question, on the other hand, considers the analysis of movements during longer time intervals, such as average movement patterns or average positions of players with respect to each other. As such, R2 can be referred to as the analysis of **team tactics**, *i.e.* the analysis of more general movement patterns that are often requested or given as guidelines by coaches in team sports. In football, for example, coaches often provide their players with a team formation as reference (e.g. 4 defenders, 4 midfielders and 2 attackers) according to which they would want their players to try and position themselves during the match. However, since for **R1** and **R2** no exact sport of interest was specified, it was to be expected that during the course of the research, new insights for possible applications of the QTC method for sports analytics would be discovered. The intention of performing research on these (at that time unknown) applications for sports analytics in the broader sense of the word is expressed by the inclusion of the third research question (**R3**).

Since QTC is an established method that was introduced prior to the research presented in this dissertation, the real research objective can be described as the finetuning and the calibration of the method to achieve the research objective mentioned above, by analysing the three research questions more specifically. The finetuning and calibration of the method, however, is not necessarily delineated by the three research questions, as these questions make use of the same method with different purposes. Consequently, the three research questions can be further specified and reformulated by following four more specific research questions[2]:

Ra **How to effectively compare different QTC-representations of sports intervals (team behaviour) or sports moments (team tactics) ? How to calculate similarities between different sports moments or intervals and decide whether two situations are similar? How can the results of the similarity calculation be validated?**

Rb **For which sports could QTC be used as a method for sports analytics?**

Rc **In what manner can QTC be enhanced to be more suitable as a method for sports analytics:**

- **What is the impact of the use of different QTC-variants?**
- **What is the impact of the use of static points?**
- **What is the impact of using different thresholds for the assignment of QTC-characters?**
- **What is the impact of the inclusion or exclusion of the positional data of the ball in the case of ball sports?**

[2] The specific research questions are labelled with letters of the alphabet. The general research question(s), which are labelled with numbers, to which the specific research question relates, to are noted in between brackets

Rd Can QTC be used for more detailed movement analysis, *i.e.* analysis of individual athletes by description and comparison of parts of the body by means of QTC?

The first specific research question, **Ra**, addresses the need for a suitable calculation of similarities between the QTC-descriptions of different movements that occur in sports. For this, multiple questions can be posed, e.g. whether it is possible to compare movements that have different temporal lengths or different numbers of players involved, how similarity between two movements can be calculated and presented by means of one similarity or distance value, what thresholds (with respect to this distance value) should be used for defining two movements as 'similar' and whether it is possible to perform the calculations in real-time, *i.e.* during an official sports event. **Rb** questions the applicability of the methodology on different sports, which will be investigated by applying the method during more concrete experiments on specific sports. The extent to which this can be investigated is highly dependable on the sports data availability and the availability of specific sport experts that can validate the results of the similarity analysis (**Ra**).

Research question **Rc** entails the different (theoretical) enhancements that could be applied to QTC in order to make the method more suitable for sports analytics. Since different variants of QTC were introduced (see Chapter 2), first it is interesting to investigate which of these variants are useful for sports analytics. Because the different variants were invented for applications in different fields, the idea is that the best QTC-variant might differ for each sport. Other enhancements are the introduction of static points, the use of different thresholds (or angular tolerances) for the assignment of QTC-characters and the inclusion of the ball. Because a good understanding of **Rc** requires the complete comprehension of QTC, these enhancements are further discussed in Sections 3.3, 3.4, 3.5 and 3.6. Research question **Rc** is in fact strongly linked with **Rb**, since the finetuning and calibration of the QTC method for sports analytics depends upon the number of sports that can be analysed during this research.

The last specific research question (**Rd**) is closely linked with **R3** and expresses the need of research on whether the esteemed method might also be usable for sports analytics on different (spatial) scales. Preliminary research on samba (Chavoshi *et al.*, 2015), for example, suggested that QTC might be useful for describing relations between body parts of individual samba dancers. It would therefore be interesting to investigate whether, and to what extent, the method that was developed for **R1** & **R2** would be applicable for movement analysis of individual athletes, by describing the relative movement of markers positioned on their body by means of QTC. However, since such data is usually obtained in an (indoor) laboratorial setting, the possible domains of the study of this research question are rather limited.

1.2.2 Research approach

To answer the four research questions that were introduced in Section 1.2.1, three consecutive research efforts are required. Firstly, programming efforts are needed to create a prototype of a program that facilitates the analysis (big) sports datasets, containing positional data, with QTC. The first months of the research, as such, were allocated for programming the prototype in the Python programming language. Advantages of the Python programming language for this research are the relative easiness for creating prototypes and easy communication with other programming languages (which are expected to be encountered in the shattered domain of sports, which lacks consensus regarding data formats, programming languages, etc.). A downside of the Python programming language, however, is the longer expected runtime when compared to other interpreted languages. The prototype of this program should include parameters to include the concepts mentioned in **Rc**. To ensure safe code storage and version control, the code of the prototype is periodically uploaded to a private Github repository (https://github.ugent.be/SportsAnalytics/QTC-sports-analytics). Moreover, Github enables colleagues to obtain the latest version of the code and, if needed, conduct and share their own programming efforts. In the development of the prototype, little efforts should be done on the optimization of the code. Indeed, the focus of the research lies in the finetuning of the methodology and the evaluation of its use for sports analytics. Secondly, sports data needs to be collected from companies, teams or even individual athletes. As these data are not shared easily, efforts for making contacts with data owners needed to start at the beginning of the research, to ensure that interesting datasets could eventually be obtained free of charge[3]. Thirdly, experiments with specific goals needed to be conducted on the obtained datasets, to extract knowledge for the finetuning of the QTC method. Finetuning can be done by comparing the results of repeated experiments using different parameter values in the prototype. To achieve this, it was essential to get in touch with sports experts who can validate these results, so that the best parameter values could be determined. Consequently, similar to the early start of the efforts on dataset collection, the search for sport experts started at early stages of the research, thus at the same time as the prototype was programmed. It must be noted that, due to learning effects, at later stages of the research, different experimental set-ups might be considered as more appropriate than the ones that were conducted at earlier stages. For this reason, it is not straightforward to compare all of the results of the different experiments. Adding to this difficulty is the fact that the experiments are conducted on different sports, making comparison a very challenging task. As research in general is a gradual process, this can be considered as unfortunate but inherent to the process of gaining insights in a new domain. Experiments are conducted on different sports, such as volleyball, basketball, golf, hockey, billiards, samba, tango, walking and running. However, only the experiments that could be considered to meet the standards of a scientific paper, thus with a clear set-up, representative dataset and rigorous (successful) validation of the results, are discussed in this dissertation. The experiments that do not meet the above-mentioned standards are omitted from this dissertation.

[3] To obtain sports datasets free of charge, it is often required to sign a non-disclosure agreement (NDA) with the party providing the data, ensuring them that the valuable dataset will not be shared with third parties or would be used for purposes other than research. Accordingly, there are two big disadvantages for conducting research on datasets that are shared under an NDA. First, the draw-up of the NDA takes up a lot of time that can therefore not be used for conducting research. Secondly, companies providing data under an NDA, often require that researchers get their explicit consent for publishing results based on the dataset. As such, at the time of experimenting, it is not sure that the permission will be granted to publish results of the conducted research.

1.3 Outline of the dissertation

This dissertation started with an introduction to the domain of sports analytics. Subsequently, the research objectives and the research approach were presented. In the remainder of this first chapter, an outline of the dissertation is given. This dissertation consists of 10 chapters. Besides the current introduction chapter, all other chapters are organised in three parts. Part I describes the used framework and consists of 4 chapters. In Part I, Chapters 2, 3 and 4 present the theoretical background of the methodology used in the research, while in Chapter 5 the implementation of the Python program is discussed. Chapter 2 presents the state-of-the-art, the latter two present the additions that are made to this method for sports analytics. The concepts that are introduced in these chapters are illustrated by means of three-cushion billiards. Three cushion billiards is chosen for its relative simplicity: it consists of three balls in a closed environment. The behaviour of the balls is, in contrast with the aspect of the human influence in other sports, only defined by the initial hit on the ball of the player with the cue stick, and further by the laws of physics. As such, three-cushion billiards is ideal to introduce the concepts in a simple way. Part II of this dissertation consists of use cases, *i.e.* practical applications of the proposed method for sports analytics. Each use case is presented as a separate chapter and, because of the different nature of sports analytics for each type of sports, limited to one sport at a time. Each of the chapters in Part II therefore includes an overview of the state-of-the-art of sports analytics for that sport and a more thorough presentation of the capabilities of the proposed method to analyse that sport or application. Part III consists of two chapters. A general discussion is included in Chapter 9, to evaluate the research questions with the results of the use cases. At last, Chapter 10 presents the general conclusion of this dissertation. Reference lists, containing all cited literature are provided separately at the end of every chapter. Following these chapters, the appendices, a summary in Dutch and the curriculum vitae of the author are included.

PART I - FRAMEWORK

Chapter 2 presents an overview of the state-of-the-art in spatio-temporal research and more specifically the Qualitative Trajectory Calculus, a qualitative spatio-temporal calculus that was introduced by Van de Weghe (2004). As this method predates the current dissertation, only the basics of the method are presented in this chapter. Challenges with respect to the application of QTC for sports analytics are presented at the end of the chapter.

Chapter 3 discusses the additions and changes that are made to the QTC calculus to be able to use it for dynamic sports analytics. First, it shows how QTC-matrix sequences can be used for storing QTC-relations between more than two players and how they can be used to describe movements during multiple time intervals. Second, the challenges that occur while using QTC for dynamic sports analytics and their methodological impacts are discussed, such as the use of static points, different threshold for the assignment of QTC-characters, the use of different QTC-variants and the use of permutations of the QTC-matrices. The impact of these challenges are discussed using simple sports examples, by means a three-cushion dataset. Since the proposed method for dynamic sports analytics requires the calculation of distances between different sports events, Chapter 3 continues with an introduction to basic methods for distance calculation methods between sequences. As these methods do not allow for the distance calculation between (sports) events with different temporal lengths, a novel method for distance calculation between (QTC-) matrix sequences of different length is presented. This method allows to compare sport events that have different temporal lengths, using QTC. Its application is demonstrated, again, by a three-cushion billiards example, by the comparison of a series of different opening shots.

Chapter 4 discusses the use of QTC as a method for static sports analytics. At first, the required additions and changes to the QTC_S-variant are discussed. This includes a new manner for the calculation of distances between the QTC_S-representations of different sport moments. Next, the challenges and solutions regarding the application of the method for static sports analytics are discussed. They are discussed rather briefly because they are very similar to the challenges for dynamic sports analytics mentioned in Chapter 3. Each challenge and its solution is illustrated by means of three-cushion billiards examples. To conclude the chapter, a more comprehensive three-cushion billiards example case is presented

Chapter 5 includes an overview of the implementation of the two methods. Two separate Python programs (one for dynamic and one for static sports analytics) are presented by means of generalized flowcharts. After that, different options for the implementation of QTC are discussed, along with their effects on real-world sports cases.

PART II – USE CASES

The second part of this dissertation consists of three use cases that are presented in separate chapters, to evaluate the proposed methods for dynamic (team behaviour) and static (team tactics) sports analytics. Each of the use cases is designed to provide evidence to answer the posed research questions. An evaluation of their value for answering the research questions is given comprehensively in Part III.

Chapter 6 contains a use case based on a real football match to evaluate the use of the dynamic football sports analytics method for player movement pattern recognition in football (team behaviour). A discussion of the state-of-the-art of movement pattern detection and recognition in sports and football is included at the beginning of the chapter. Next, the dataset and the set-up of the use case are described, after which the results are presented. To test the significance of the results, a statistical validation of the results, including a football expert panel is included.

Chapter 7 consists of a use case aimed at investigating the use of the proposed method for dynamic sports analytics for a more detailed analysis of movements of the human body. For this, the use case includes an analysis of human walking behaviour, a domain which is called gait analysis. First, a literature overview of relevant research on gait analysis is provided. After that, the dataset and use case set-up are discussed. Third, a methodology of the statistical validation of the results is explained. Next, the results are presented and validated. To conclude, the significance and value of the results for the domain of gait analysis are discussed.

Chapter 8 presents a use case that is aimed at evaluating the use of the static sports analytics method for analysing team tactics. More specifically, the proposed method is applied on team formation analysis in football. First, a literature overview provides the reader with the necessary background information on team formation analysis for ball sports in general and for football more specifically. The use case consists of a series of basic football examples and one more exhaustive football experiment. The third section of the chapter therefore introduces the used datasets and analyses set-up. Subsequently, the basic examples of team formation analysis in football are presented, followed by the football experiment. It includes a validation of the results of the football experiment based on match reports published by popular sports media. The chapter is concluded by a discussion of the value of the static sports analytics method for team formation analysis in football, based on the results of the examples and the experiment.

PART III – GENERAL DISCUSSION AND CONCLUSION

Chapter 9 recapitulates the posed research questions and presents a summary of the findings based on the results of the use cases of Part II of this dissertation. Besides a summary, a thorough discussion of the answers on the research questions the research questions is included. A discussion of the use cases' results for their respective specific domains is not presented in this chapter, as this is included in the chapters of Part II of this dissertation. To end the chapter, some suggestions of possible avenues for future research are presented.

Chapter 10 is a brief chapter that presents the general conclusion of this dissertation.

REFERENCES

Almar, C. B. (2013). Sports Analytics: A guide for coaches, managers and decision makers. New York, NY, U.S.A.: Columbia University Press.

Benkraiem, R., Louhichi, W., Marques, P. (2009). Market reaction to sporting results: the case of European listed football clubs. *Management Decision*, 47(1), 100-109.

Bialkowski, A., Lucey, P., Carr, P., Yue, Y., Matthews, I. (2014). Win at Home and Draw away: Automatic Formation Analysis – Highlighting the Differences in Home and Away Team Behaviors. *Disney Research,* Pittsburgh, PA, U.S.A.

Bonomo, F., Duran, G., Marceno, J. (2014). Mathematical programming as a tool for virtual soccer coaches: a case study of a fantasy sport game. *International Transactions in Operational Research*, 21(3), 399-414.

Chavoshi, S.H., De Baets, B., Neutens T., De Tre, G., Van de Weghe, N. (2015). Exploring dance movement data using sequence alignment methods. *PLOS One,* 10(7), 1-25.

Coleman, B. J. (2012). Identifying the "Players" in sports analytics research. *Interfaces*, 42(2), 109-118.

Cooper, A. (2012). What is analytics? Definition and essential characteristics. *CETIS analytics series,* 1(5), 1-10.

Cristani, M., Cohn, A.G., Benett, B. (2000). Spatial locations via morpho-mereology. In A.G. Cohn, F. Giunchiglia, B. Selman (Eds.), Proceedings of the 7th conference on principles of knowledge representation and reasoning (pp. 15-25). Breckenridge, CO, U.S.A: Morgan Kaufman

Delafontaine, M., Cohn, A.G., Van de Weghe, N. (2011). Implementing a qualitative calculus to analyse moving point objects. *Expert Systems with Applications*, 38(5), 5187-5196.

De Luca, M., Horovitz, R., Pitt, B. (Producers), & Miller, A. (Director). (2011). *Moneyball* [Motion Picture]. United States: Colombia Pictures.

D'Orazio, T., Leo, M. (2010). A review of vision-based systems for soccer video analysis. *Pattern Recognition*, 43(8), 2911-2926.

Duarte, R., Araujo, R., Folgado, H., Esteves, P., Marques, P., Davids, K. (2012). Capturing complex, non-linear team behaviours during competitive football performance. *Journal of Systems Science and Complexity*, 26(1), 62-72.

Frank, A.U. (1996). Qualitative spatial reasoning: cardinal directions as example. *International Journal of Geographical Information Science*, 10(3), 269-290.

Franks, I. Hughes, M. (2004). Notational analysis of sport: Systems for better coaching and performance. London, U.K.: Routledge.

Freksa, C. (1992). Using orientation information for qualitative spatial reasoning. *Lecture Notes in Computer Science*, 639, 162-178.

Fung, S.K., Sundaraj, K., Ahamed, N.U., Kiang, L.C., Nadarajah, S., Sahayadhas, A., Ali, M.A., Islam, M.A., Palaniappan, R. (2014). Hybrid markerless tracking of complex articulated motion in golf swings. *Journal of Bodywork and Movement Therapies*, 18(2), 220-227.

Goodhart, C. (1981). Problems of monetary management: the U.K. experience. In A.S. Courakis (Ed), Inflation, Depression, and Economic Policy in the West (pp. 111-146). Lanham, MD, U.S.A.: Rowman & Littlefield.

He-zhou, L.I. (2006). The study of the trend of badminton rules through rules evolution of net sports events. *Journal of Guangzhou sport university,* 1, 016.

Hughes, M. (2007). How to choose your table tennis balls. [online] [viewed 27/05/2019]. Available from: https://www.allabouttabletennis.com/table-tennis-balls.html

Ingle, S. (2015). How Midtjylland took the analytical route towards the champions league. [online] [viewed 10/12/2018]. Available from: https://www.theguardian.com/football/2015/jul/27/how-fc-midtjylland-analytical-route-champions-league-brentford-matthew-benham

Iwasaki, Y. (1997). Real-world applications of qualitative reasoning: Introduction to the special issue. *IEEE Intelligent System*, 12(3), 16-21.

Jäger, J.M., Schöllhorn, W.I. (2007). Situation-orientated recognition of tactical patterns in volleyball. *Journal of Sports Sciences*, 25(12), 1345-1353.

Keane, T. (2018). New data regulations will be a game-changer for sport. [online] [viewed 02/02/2019]. Available from: https://www.independent.ie/sport/new-data-regulations-will-be-a-gamechanger-for-sport-36741211.html

Lengel, D. (2018). Has baseball analytics killed the art of hitting? [online] [viewed 7/03/2019]. Available from: https://www.theguardian.com/sport/2018/oct/02/has-baseball-analytics-killed-the-art-of-hitting

Levinson, D., Christensen, K. (1996). Encyclopedia of world sport. New York, NY, U.S.A.: Oxford University Press.

Link, D. (2018). Sports Analytics. How (commercial) sports data create new opportunities for sports science. *German Journal of Exercise and Sports Research,* 48(1), 13-25.

Liponski, W. (2003). World sports encyclopedia. St Paul, MN, U.S.A.: MBI.

Lucey, P., Bialkowski, A., Carr, P., Yue., Y., Matthews, I. (2014). How to get an Open Shot: Analyzing Team Movement in Basketball using Tracking Data. *Disney Research,* Pittsburgh, PA, U.S.A.

Marcon, G., Franchini, E., Jardim, J., Neto, T.L.B (2010). Structural analysis of action and time in sports: judo. *Journal of Quantitative Analysis in Sports,* 6(4), 1-13.

McCaskill, S. (2018). Data analytics makes tennis players and umpires better but will it destroy human element? [online] [viewed 7/03/2019]. Available from: https://www.forbes.com/sites/stevemccaskill/2018/11/20/data-analytics-makes-tennis-players-and-umpires-better-but-will-it-destroy-human-element/#320618352a79

Mohamed, T. (2019). Juventus' shares are soaring after Cristiano Ronaldo's hat-trick took the team into the final 8 of the Champions League. [online] [viewed 13/05/2019]. Available from: https://markets.businessinsider.com/news/stocks/juventuss-shares-are-soaring-after-cristiano-ronaldos-hat-trick-2019-3-1028027048

Morgulev, E., Azar, O.H., Lidor, R. (2017). Sports analytics and the big-data era. *International Journal of Data Science and Analytics,* 1, 1-10.

O'Donoghue, P. (2015). An introduction to performance analysis of sport. London, U.K.: Routledge.

Perin, C., Vuillemot, P., Fekete, J.D. (2013). Soccerstories: A kick-off for Visual Soccer Analysis. *IEEE Transactions on Visualization and Computer Graphics*, 19(12), 2506-2515.

Reportstack (2015). New report shares details about the sports analytics market size at $125 million in 2014 and is anticipated to reach $4.7 billion by 2012. [online] [viewed 6/11/2018]. Available from: https://www.whatech.com/market-research/consumer/archive/72109-new-report-shares-details-about-the-sports-analytics-market-size-at-125-million-in-2014-and-is-anticipated-to-reach-4-7-billion-by-2021

Reunes, T. (2014). Analyse van bewegingstrajecten ter optimalisatie van voetbaltactiek. Unpublished scription: Universiteit Gent, Vakgroep Telecommunicatie en Informatieverwerking.

Schwarz, E.C., Hunter, J.D. (2012). Advanced theory and practice in sport marketing. London, U.K.: Routledge.

Seweryniak, T., Mroczek, D., Lukasik, L. (2013). Analysis and evaluation of defensive team strategies in women's beach volleyball – an efficiency-based approach. *Human Movement*, 14(1), 48-55.

Shmueli, G. (2010). To explain or to predict? *Statistical Science,* 25(3), 289-310.

Siegel, E. (2013). Predictive analysis: The power to predict who will click, buy, lie or die. Hoboken, NY, U.S.A.: Wiley and Sons.

Skinner, B., Guy, S.J. (2015). A method for using player tracking data in basketball to learn skills and predict team performance. *PLOS One,* 10(9), 1-15.

Snijders, C., Matzat., U., Reips, U.D. (2012). 'Big Data': Big gaps of knowledge in the field of internet science. *International Journal of Internet Science,* 7(1), 1-5.

Squires, D. (2016). The illustrated history of football. London, U.K.: Cornerstone.

Stockwell, S. (2015). Using live data and social amplification to drive fan engagement. *Journal of*

Direct, Data and Digital Marketing Practice, 16(4), 275-277.

Swartz, T. B. (2018). Where should I publish my sports paper? *The American statistician*, 1(0), 1-18.

Timms, A. (2018). Are super-nerds really ruining US sports? [online] [viewed 7/02/2019]. Available from: https://www.theguardian.com/sport/2018/aug/16/sports-nerds-analytics-data

Van Belleghem, K., Denecker, M., De Schreye, D. (1994). Representing continuous change in the abductive event calculus. In P. Van Hentenrijck (Ed.), Proceedings of the international conference on logic programming (pp. 225-240). Camebridge, MA, U.S.A.: MIT Press.

Van de Weghe, N. (2004). *Representing and Reasoning about Moving Objects: A Qualitative Approach*. Ph.D. dissertation, Ghent University.

Van de Weghe, N., Kuijpers, B., Bogaert, P., De Maeyer, P. (2005). A qualitative trajectory calculus and the composition of its relations. In M.A. Rodriguez, I. Cruz, S. Levashkin, M.J. Egenhofer (Eds.), Proceedings of Geospatial Semantics (pp. 60-76). Berlin, Heidelberg, Germany: Springer.

Waller, M.A., Fawcett, S.E. (2013). Data science, predictive analytics, and big data: A revolution that will transform supply chain and design of management. *Journal of Business Logistics,* 34(2), 77-84.

Whannel, G. (2005). Media sport stars: Masculinities and moralities. London, U.K.: Routledge.

Wright, M. B. (2009). 50 years of OR in sport. *Journal of Operational Research Society,* 60(1), 161-168.

Yoshida, M., Gordon, B., Nakazawa, M., Biscaia, R. (2014). Conceptualization and measurement of fan engagement: Empirical evidence from a professional sport context. *Journal of Sport Management,* 28(4), 399-417.

PART I – FRAMEWORK

2.

THE QUALITATIVE TRAJECTORY CALCULUS

Parts of this chapter were modified from: BEERNAERTS, J., DE BEVER, E., LENOIR, M., DE BAETS , B., VAN DE WEGHE, N. 2019. A method based on the Levenshtein distance metric for the comparison of multiple movement patterns described by matrix sequences of different length. *Expert Systems with Applications,* 115, 373-385.

The methodology used in this research is based on the Qualitative Trajectory Calculus (QTC; Van de Weghe, 2004), a qualitative spatio-temporal calculus that originated from the field of geography. QTC describes the relative motion of disconnected moving objects, thus providing an answer to many trajectory-related questions. The theoretical framework of QTC has been well studied and documented by composition tables (Van de Weghe *et al.,* 2006) and conceptual neighbourhood diagrams (Van de Weghe & De Maeyer, 2005). Several variants of QTC have been described (Van de Weghe *et al.*, 2005), which will be presented in Section 2.2. QTC makes abstractions of real-world situations using four simplifications, which are shown in Figure 2-1 (Delafontaine *et al.*, 2012). Although in Figure 2-1 the moving objects are footballers, there is no requirement for the moving objects to be individual persons of objects. The movement of any object or point of which the geographical location can be tracked through time can be described by QTC. The first of the four simplifications is the relational simplification, which means that only the relation between exactly two Moving Objects (MOs) is interpreted. Second, by an object simplification, the moving objects are simplified to Moving Point Objects (MPOs). This is done by assuming the considered object is a 1-dimensional point rather than an object with more complex spatial dimensions (such as footballers on a football field, who have arms, legs, etc.) Simplification can be done in various ways, for example by using the midpoint of the relevant 2-dimensional representation of the MO, the 2-dimensional representation of its centre of mass, or by choosing a point of the MO that is of special interest to the researcher, such as the midpoint between the two feet of the footballers shown in Figure 2-1. The third simplification is the restriction to disjoint MPOs, the so-called topological simplification. Fourth, the temporal simplification implies that QTC-relations are relations that hold during a particular time interval. All of the four simplifications are used for all of the variants of QTC.

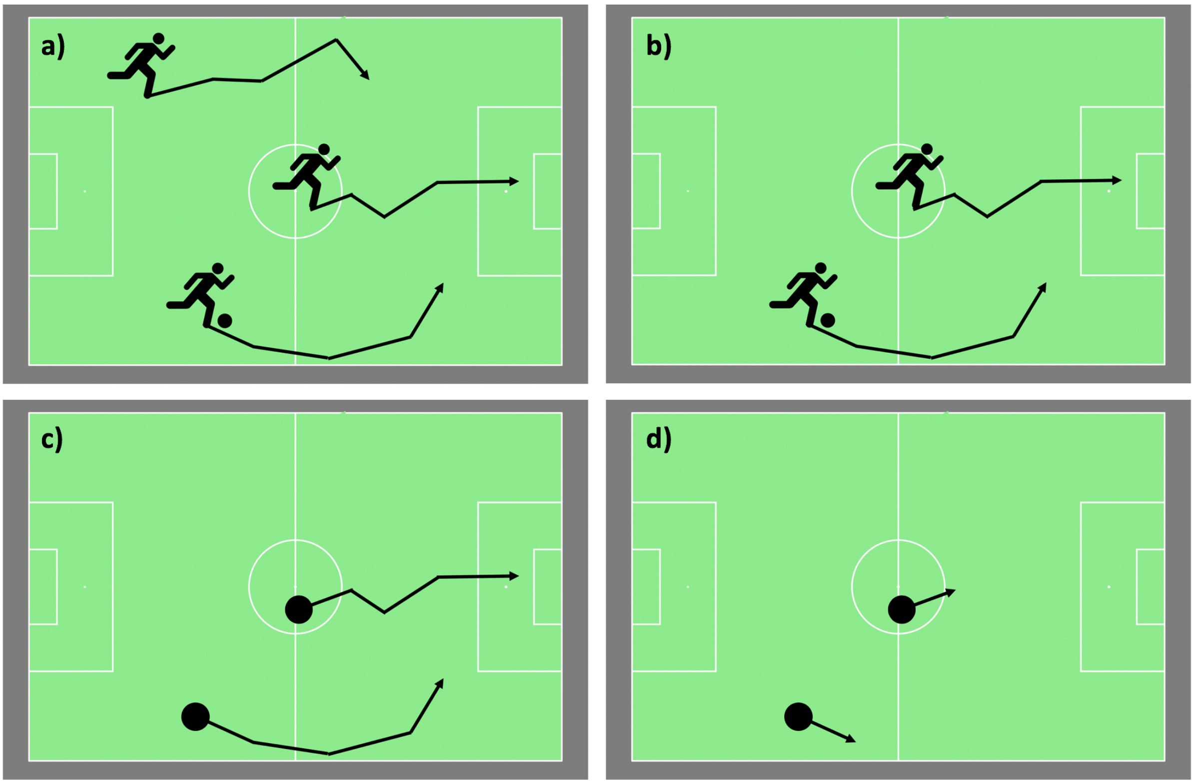


Fig. 2-1. Simplifications of QTC shown on a real-world (football) example. The trajectories of three football players and their movement patterns on a football field (a), which are consecutively simplified by the relational simplification (b), the object simplification (c) and the topological simplification (d).

The first section of this chapter gives an introduction to the field of spatio-temporal reasoning. After this, the different QTC-variants are discussed in separate sections. As those sections mainly present research that was conducted previously, the aim is not to present an exhaustive overview but rather sufficient background information for the reader to be able to easily understand the following chapters of this dissertation. To conclude the chapter, some challenges of analysing sports data with QTC are discussed.

2.1 Spatio-temporal reasoning

QTC originates from the field of spatio-temporal reasoning about moving objects, a field combining temporal and spatial reasoning. Looking at temporal reasoning, an important work is the overview Allen (1981) made of how to reason with temporal knowledge, based on an interval-based representation. This influential work was further enhanced by Galton (1990) making it suitable for assessing continuous changing temporal phenomena. For this, Galton introduced a series of modifications to Allen's work including the broadening of the original temporal ontology to include time instants on the same level as the time intervals (Galton, 1990). More recent work on temporal reasoning and the 2D representations of time intervals was conducted by Kulpa (1997, 2006), who introduced the Triangular Network (TM). Zhang *et al.* (2016) later extended this method to the Continuous Triangular Model (CTM), by including a third dimension representing the values of an attribute during certain time intervals. In their work they exemplified the method with a sports analytics example, by visually representing variables such as speed, velocity, ball possession of football players during a real football match.

During the late 20th century as well as in more recent years, a lot of research has been done in the field of spatial reasoning. Qualitative spatial reasoning in particular has seen a huge growth (Cohn, 2005), thereby enabling detailed analysis on big datasets with limited computation power. One of the reference works for describing qualitative relations between objects is the topological relations framework of Egenhofer (Egenhofer, 1989). As it was later extended to include geographical regions (Egenhofer & Herring, 1991), relations between fuzzy area objects (Dijkmeijer & de Hoop, 1996), three-dimensional relations (Billen & Zlatanova, 2003) and region-region relations (Egenhofer, 2010), it is an excellent example of the emerging need for qualitative spatial reasoning. Research on topological relations such as that of Egenhofer (Egenhofer &

Shariff, 1998) was taken further by Bera and Claramunt (2003), with the introduction of topology-based measures that describe nearness relationships between regions in a spatial system. A similar framework of qualitative spatial reasoning, based on the logic of regions, is the Region Connections Calculus (RCC), incorporating complex concepts such as regions with uncertain boundaries (Randell *et al*, 1992). RCC, which describes relations between regions, consists of eight possible relations in its most common form (RCC8). Worboys (1994) however, stated that spatial relations often have temporal relationships, and that rather than posed in a subordinate role, this temporal information should be incorporated in spatial models, as a unified model of spatio-temporal relations. When studying spatio-temporal relations, the distinction between continuous and discontinuous movement can be made (Galton, 2000). In the case of continuous movement, in any sufficiently small temporal neighbourhood, any movement attribute having ordered values (such as speed, (x,y,t)-coordinates or height) will pass through all of the intermediate attribute values when changing from one value to another. In the case of discontinuous movement, however, this is not the case.

The interest in qualitative spatio-temporal reasoning about moving objects has seen a huge growth in the last decade, with considerable efforts in the formalisation of motion. Examples of this can be found in Muller (2002), Ibrahim (2007), and Kurata & Egenhofer (2009). A well-known spatio-temporal calculus is Relative Motion (REMO), which was introduced by Laube *et al.* in 2004, to describe and compare relative movements of multiple objects in a qualitative manner. For analysing spatio-temporal interaction between pairs of moving objects, Long *et al.* (2013) introduced the 'dynamic interactions' method. More recently, temporal networks were enhanced to include spatio-temporal relations (Williams & Musolesi., 2016; Von Landesberger *et al.*, 2016). In 2018, Zhang *et al.* combined the CTM with the multi-temporal scale spatio-temporal network, for exploring dynamic interactions in movement data. Just as with REMO, the method was applied on a football match to exemplify its applicability for analysing real-world situation. In spatio-temporal reasoning, matrices are often (Cohn *et al.*, 1997; Van de Weghe *et al.*, 2006) used to store spatio-temporal information on relations between objects. If the evolutions of these relations are considered, these matrices can be placed into a sequence, taking into account the changes over time. This is also the case for the

QTC, the qualitative framework for reasoning about moving objects that was adopted for this research.

There are numerous advantages of using qualitative approaches for spatio-temporal reasoning. Human spatial reasoning, for example, mostly relies on qualitative abstractions of the (too precise) quantitative space (Hazarika and Cohn, 2011). Indeed, Guan and Duckham (2011) argue that the human perspective of motion is primarily qualitative in nature. Earlier, Kuipers (1986) already stated that there is a consensus that qualitative approaches are more suitable for the representation of spatio-temporal human cognition than quantitative approaches. Moreover, qualitative representations are typically more compact and computationally efficient than quantitative methods (Dodge *et al.*, 2012). For these reasons, and as argued in the second paragraph of Section 1.2.1, the goal of this research was to investigate the use of QTC, a qualitative spatio-temporal calculus, for sports analytics.

2.2 QTC-Variants

Throughout the years, different types of QTC were developed. Different variants can be distinguished, each with different levels of detail and spatial dimensions, and are named accordingly by placing the initial of the variant subscript of the abbreviation 'QTC', followed by the number of spatial dimensions and the level of detail[4]:

$$QTC_{Variant\ initial-Spatial\ dimension-Level}$$

QTC Basic (QTC_B), the most basic QTC-variant that only takes distance constraints into account, is discussed in Section 2.2.1. In the subsequent section, QTC double cross (QTC_C) is introduced, which provides a more detailed description of the movements of

[4]Note that for the remainder of this dissertation, we will refer to the level of detail with 'level', for reasons of simplicity.

MPOs by taking the movement directions into account. This chapter introduces QTC_B and QTC_C in 1D and 2D, yet both were later on extended to the three dimensional world (Mavridis *et al.*, 2015). For reasons of completeness, the last sections discuss QTC shape and QTC network, two QTC-variants that can be found in literature but were not used in the research. As such, they are discussed rather briefly.

2.2.1 QTC Basic

QTC Basic or QTC_B is the most basic and limited variant of QTC and was the first variant to be introduced in 2004 (Van de Weghe, 2004). In a 1-dimensional environment, QTC_{B1} describes the relation between two MPOs which are moving on a 2-dimensional line (e.g. two cars driving on a single lane). These relations can be described by the level one QTC_{B11} or the level two QTC_{B12}. When considering 2-dimensional movements in space, QTC_{B21} and QTC_{B22} can be distinguished for level one and level two respectively.

For level one QTC_B, two characters describe the relative movement of two MPOs, for level two QTC_B, a third character is added to this. Furthermore, the two characters of the level one QTC_B description are also being used in the QTC Double Cross description of two MPOs, but additional characters are added.

The level one QTC_B-relation of two MPOs k and l during a certain time interval T is a tuple (a, b) consisting of two qualitative symbols that share the threefold domain D = {-, 0, +}. Character a is the movement of k with respect to l during interval T, whereas b is the movement of l with respect to k during the same interval T. The movement of k with respect to l during interval I (and vice versa) is defined as follows:

"-" if k moves towards l
"+" if k moves away from l
"0" if k is stable with reference to l (all other cases)

Depending on the spatial dimension of the QTC_B that is being considered, different numbers of combinations are possible. For level one QTC_{B11} and level one QTC_{B21}, the combination of three possibilities for two characters results in nine (3^2) possible QTC-relations. Figure 2-2 shows all the QTC_{B11}-relations, where *k* is always on the left side, and *l* on the right side. A filled dot indicates that an MPO could be stationary, whereas an open dot indicates that it must be moving along the dashed line. The fact that a dashed line is used rather than a full one, is because the relative speed is ignored for the level one QTC_{B11}. Figure 2-3 represents all of the QTC_{B21}-relations where, similar to Figure 2-2, *k* is always on the left side, and *l* on the right side. Here also, a filled dot indicates that an MPO could be stationary, whereas an open dot indicates that it must be moving. The line segments and crescents represent the potential motion areas. Note that their boundaries are open, and that, for the crescents, the straight boundaries correspond to elements of another relation. So when a crescent is used, the movement starts in the dot and ends somewhere on the curved side of the crescent and not on its straight part. Dashed lines represent uncertain boundaries caused by ignoring relative speed.

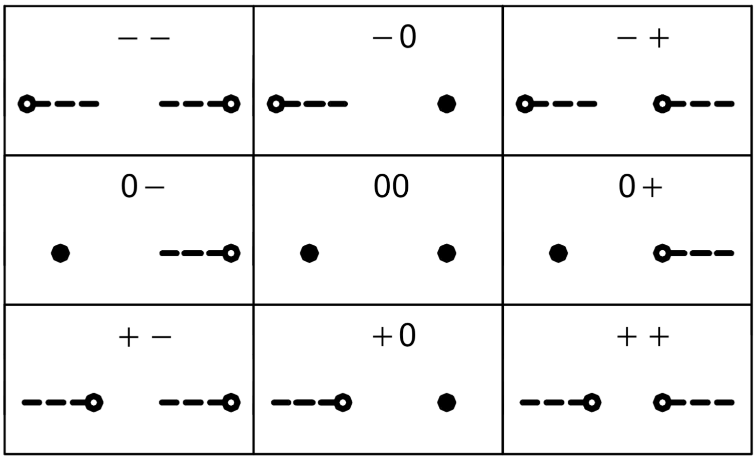
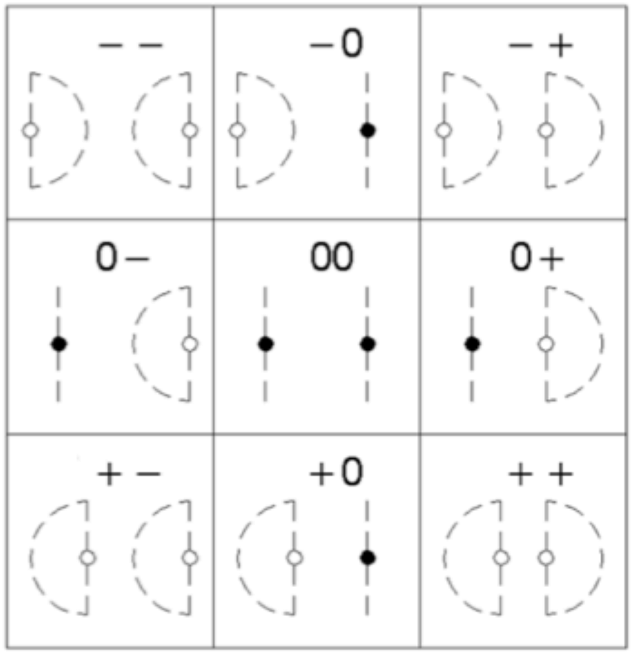


Fig. 2-2. QTC_{B11}-relations icons. (after Van de Weghe, 2004)



The level two QTC_B-relation of two MPOs *k* and *l* during a certain time interval *T* is a tuple (*a*, *b*, *c*) with characters *a* and *b* identical to the level one QTC_B description and character *c* consisting of qualitative relation symbols that share the threefold domain D = {-, 0, +}. Character *c* denotes the relative distance of the displacement of *k* during *T* with respect to the distance of the displacement of *l* during *T*. Character *c* as such is defined as follows:

"-" if the displacement of *k* is smaller than the displacement of *l* during *T*
"+" if the displacement of *k* is larger than the displacement of *l* during *T*
"0" if the displacement of *k* is identical to the displacement of *l* during *T*

As such, QTC_{B12} and QTC_{B22} describe the relative movements with a finer granularity by taking into account the relative speed (or displacement) of the MPOs. Therefore, by definition, a larger number of relations are possible. For second level QTC_B there are 27 (3^3) theoretical possibilities, although, since some combinations of characters in QTC_{B12} are impossible in the real-world, only 17 relations remain for the one dimensional case (see Figure 2-4).

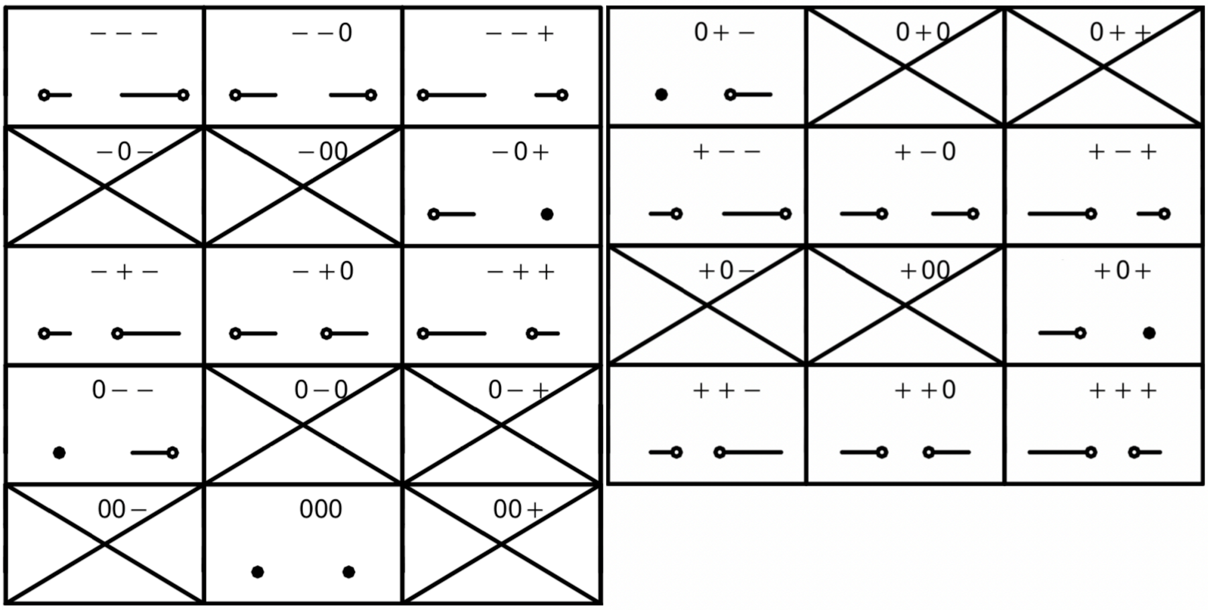


Fig. 2-4. Real-world QTC_{B12}-relations (after Van de Weghe, 2004). Cells with theoretical relations that are not possible in the real-world are crossed out.

2.2.2 QTC Double Cross

QTC Double Cross (QTC_C) describes relative movements with a finer granularity, by adding information regarding the orientation of the relative movements to the description of QTC_B, that only takes distance information into account. QTC_C has only been described in 2D (Van de Weghe, 2004) and is based on the Double Cross calculus that was introduced by Freksa (1992) and Freksa & Zimmermann (1992). In the 2-dimensional space, two levels are being distinguished, *i.e.* the level one QTC_{C21} and the level two QTC_{C22}, which are composed of respectively four and six characters.

The level one QTC_{C21}-relation of two MPOs, k and l, during a certain time interval T is a tuple (a, b, c, d) consisting of four qualitative symbols that share the threefold domain D = {-, 0, +}. Similar to level one QTC_B, character a describes the movement of k with respect to l during interval T with respect to the first perpendicular reference line ($RL\perp1$) and b describes the movement of l with respect to k during the same interval T with respect to the second perpendicular reference line ($RL\perp2$). The reference lines originate from the Double-Cross calculus (Freksa, 1992) and their application for QTC_C are shown in Figure 2-5. Similar to QTC_B, the movement of *k* with respect to *l* during interval *I* (and vice versa) is defined as follows:

"-" if *k* moves towards *l*
"+" if *k* moves away from *l*
"0" if *k* is stable in relation to *l* (all other cases)

Character *c* is the movement of *k* with respect to *l* during interval *T* with respect to the reference line (*RL*) between both MPOs, and *d* is the movement of *l* with respect to *k* during the same interval *T* with respect to the same reference line (*RL*). The movement of *k* with respect to *l* during interval *T* (and vice versa) is defined as follows[5]:

"-" if *k* is moving to the left side of *RL*
"+" if *k* moving to the right side of *RL*
"0" if *k* is moving along *RL*

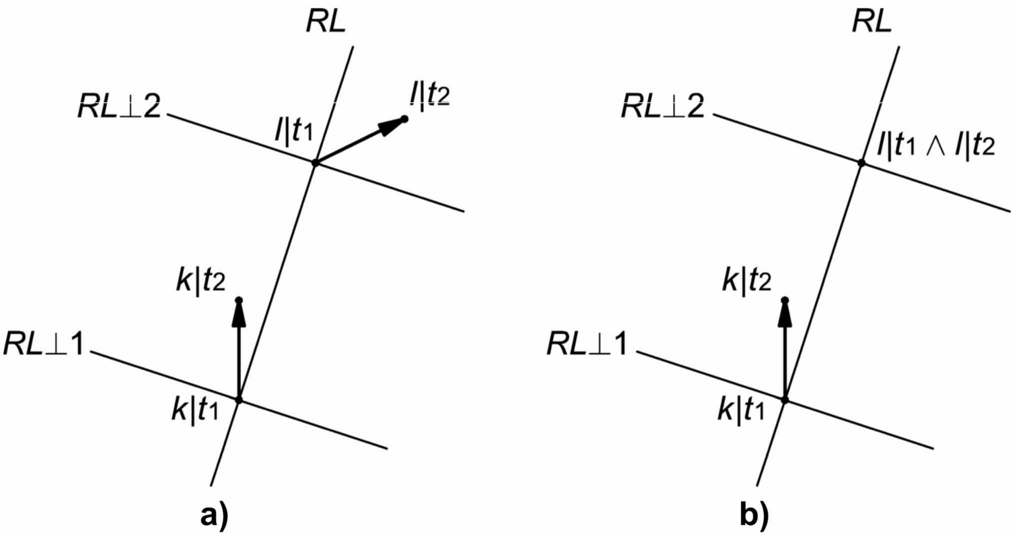


Fig. 2-5. The double cross concept for QTC with two moving objects (a) and one moving object (b). (after Van de Weghe *et al.*, 2005)

[5] Note that for determining whether the MPO is moving to the left of to the right side of the *RL*, the observer should be located at *k* and directed towards *l* (and vice versa for object *l*)

The level two QTC_{C22} adds two characters to the QTC_{C12} description and can thus be represented by a tuple (*a*, *b*, *c*, *d*, *e*, *f*), where the fifth character *e* represents the relative speed and the sixth character *f* gives a qualitative measure for the relative direction of the velocity vector.

Because, in comparison with QTC_B, more information is included into the QTC_C description, much more QTC-relations are possible. For the level one QTC_C, the combination of the threefold domain D = {-, 0, +} for the different characters result in 81 (3^4) theoretical QTC-relations which are shown in Figure 2-6. In this figure, *k* is shown as the dot on the left side, whereas *l* is the dot on the right side. A filled dot indicates that an MPO might be stationary, whereas an open dot means that it must be moving. Some icons of the relations contain disk quarters. Open disk quarters mean that the endpoint of the motion vector in these cases can be every point on the curved part of the disk quarter excluding the horizontal and vertical line segment. The dashed lines represent uncertain boundaries caused by ignoring relative speed. Note that all of the relations shown in Figure 2-6 are possible to occur for real-world 2-dimensional movements.

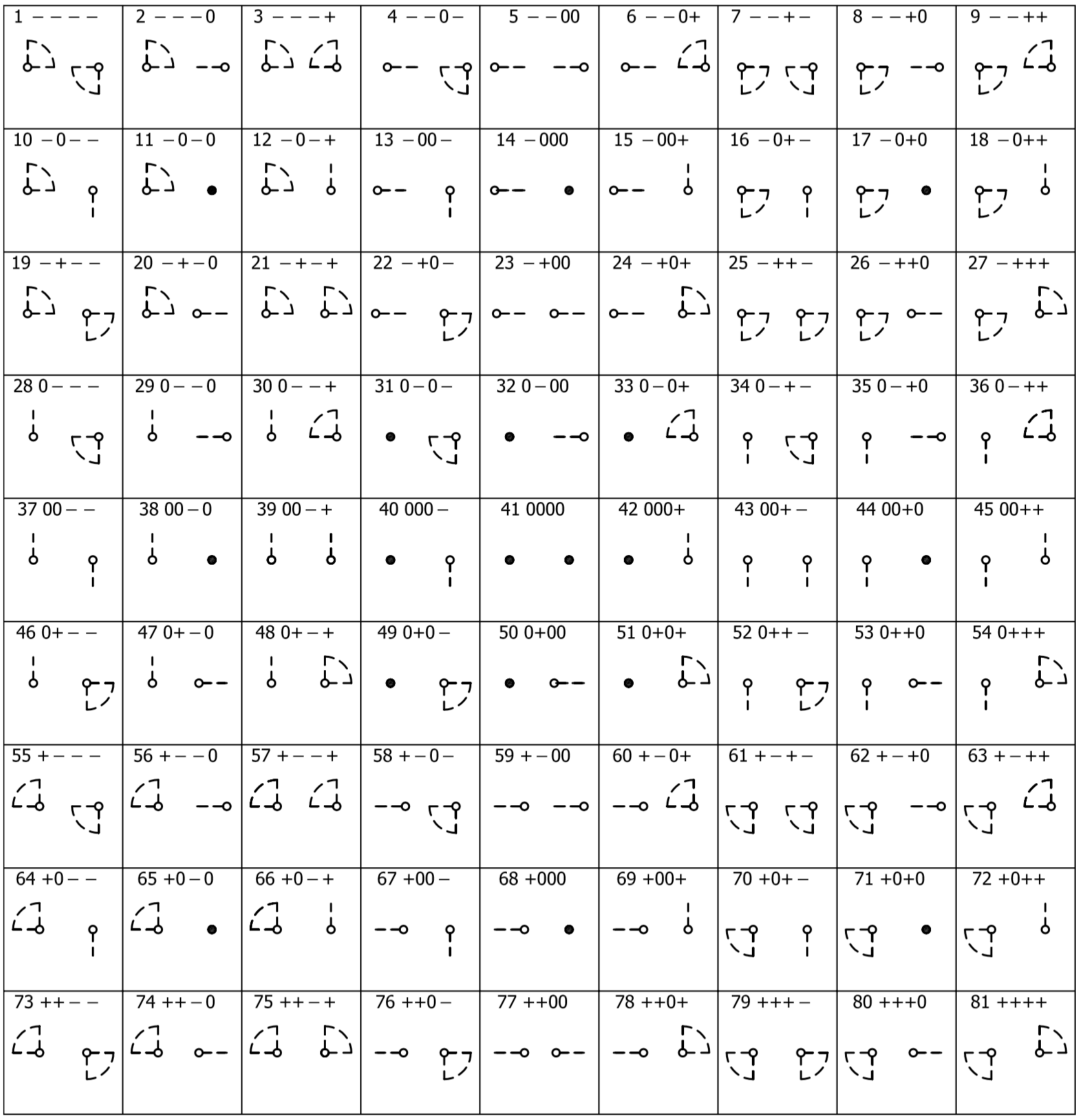


Fig. 2-6. Overview of the different possible QTC_{21}-relations. (Van de Weghe *et al.*, 2005)

For the level two QTC_C, the combination of the threefold domain D = {-, 0, +} for the different characters results in 729 (3^6) theoretical QTC-relations, of which 305 exist for real-world.

2.2.3 QTC Network

As Moreira *et al.* (1999) argue, not all MPOs move completely free in space. Often, MPOs move through space with certain constraints. In QTC_{B1}, for example, MPOs are bound to follow a one-dimensional linear path. In the real-world, objects that can be represented by MPOs (such as cars, trains or even humans), often move along networks such as roads, railways, etc.. Topological and directional calculi (such as QTC_B and QTC_C), however, are not very well suited for describing relative movements of MPOs on a network (Bogaert *et al.*, 2007) since they do not take the spatial structure of a network into account. Bogaert *et al.* (2007) therefore introduce QTC Network (QTC_N) to qualitatively reason about the relative movement between two MPOs which are constrained in their movement by a network. While QTC_N could be used for analysing certain sports, e.g. sailing and cycling, it was not used in this dissertation. The reason for this is that most of the sports for which datasets were available are characterized by movement that is not bound to networks (e.g. football, basketball and golf).

2.2.4 QTC Shape

The idea of using QTC for describing shapes dates back to the introduction of the calculus in 2004 (Van de Weghe, 2004) and was formalised by Van de Weghe, De Tré, Kuijpers & De Maeyer in 2005 with the introduction of QTC_S (QTC Shape). While other QTC-variants describe relative movement, the lack of movement when describing shapes is compensated by the construction of vectors between the points that define the shape. These vectors as such are not considered as representations of movements between two time points but rather as vectors at a single time moment. Consequently, the relations between each of the vectors can be described in a manner similar to QTC_B or QTC_C.

While determining the defining points of a shape can be done relatively easy for polygons, the description of shapes with curved edges with QTC_S requires the use of the generalization concept (Van de Weghe, De Tré, Kuijpers & De Maeyer, 2005). Since real-world applications of QTC_S were not yet described in the literature, we opt to further introduce QTC_S and elaborate on its use for sports analytics in Chapter 4.

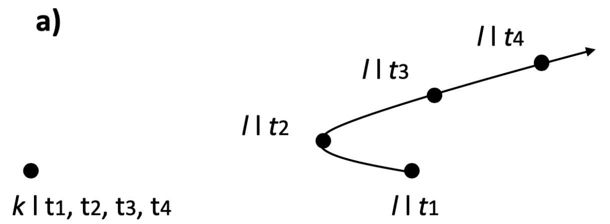
2.3 Challenges of applying QTC on sports

Using QTC for sports analytics is not a straightforward task. A series of challenges regarding the calculus as well as some practical issues need to be solved. In this section these challenges and issues are introduced by means of general examples. Similar to the division between the QTC-variants describing movement (QTC_B, QTC_C) and the variant describing shapes (QTC_S) in Section 2.2, in this section the division is made between QTC for dynamic sports analysis and QTC for static sports analysis. The method for dynamic sports analysis can be used to analyse team behaviour (**R1**), while the analysis of team tactics (**R2**) can typically be seen as a static sports analysis, using QTC_S. For investigating research question three, both the dynamic as well as the static sports analytic methods can be used.

2.3.1 Dynamic sports analytics

Dynamic sports analytics entail the analysis of movements in sports and, in this research, makes use of the QTC-variants describing movement (QTC_B & QTC_C). Applied on sports, the dynamic movement analysis can be described as an analysis of team behaviour. Movements in sports are typically continuous movements (Galton, 2000), *i.e.* a moving player passes through all intermediate points between his begin and end position on a field. To analyse these movements with QTC, however, it is required that sports data needs to be collected. The problem is that, although modern data collection technologies allow for coordinate logging with high spatial accuracy and temporal frequency (Morgulev *et al.*, 2017), only data at discrete timestamps can be collected. Consequently, the theoretical continuous character of QTC might be violated when analysing sports data, as shown in Figure 2-7. On the left of this figure, a (temporary) stationary MPO k and a moving MPO l are shown, of which the latter moves along a linear trajectory. On the right of the figure, the QTC_{B21}-relations between the two MPOs are shown, for the case where their locations are logged at four discrete timestamps $t1$, $t2$, $t3$, $t4$. The QTC_{B21}-relation of l with respect to k directly switches from a '-' to a '+'

when going from time interval T(1-2)[6] to time-interval T(2-3), without first becoming a '0'. In the case of a continuous system, however, it is always possible to find a timestamp between the time intervals T(1-2) and T(2-3) where the QTC_{B21}-relation of l with respect to k is '0'. The example in Figure 2-7 thus illustrates that in practice the continuous character of QTC is not guaranteed. Assuming that sports data are collected with a temporal resolution that is fine enough to detect all important movement interactions between the MPOs, we ignore this issue for the sports analytics based on QTC.



b)

Time interval	QTC_{B21} -relation
T(1- 2)	0 -
T(2- 3)	0 +
T(3- 4)	0 +

Fig. 2-7. The location of two MPOs k and l logged at four timestamps $t1$, $t2$, $t3$ and $t4$ (a) and their respective QTC_{B21}-relations (b).

Furthermore, when describing the relative movements of multiple MPOs during multiple time intervals, the QTC-relations need to be stored in a different manner than presented in Figure 2-7, to ensure readability and enable easy comparison of two sets of movements. In literature, the idea of storing these relations in QTC-matrices and QTC-matrix sequences is described by Van de Weghe (2004), but it was never tested whether this is achievable in practice for larger datasets. Consequently, the question then arises on whether and how it is possible to compare movements that contain different numbers of MPOs and/or different temporal lengths. For this, a new method for the distance calculation between matrix sequences of different lengths needs to be introduced, which will be presented in Chapter 3 of this dissertation.

Another challenge is the differentiation between actions occurring on different absolute locations. Since QTC describes relative movements between MPOs, two identical movements happening at different locations, e.g. on the defensive or offensive half of a

[6] In this dissertation, time intervals will be noted as T(begintimestamp-endtimestamp)

football field, are expected to have identical QTC-descriptions (and consequently a distance value of zero between them). The idea is therefore to include static points that can differentiate between these two movements, into their respective QTC-descriptions. By including such static points, the QTC-descriptions of these two movements will at least have one different QTC-character, resulting in a distance value that is > 0. Which static points need to be included can be investigated separately for each different purpose, and will be presented in different case studies in this dissertation. The challenge of applying static points will be discussed more in detail in Chapter 3.

A different challenge lies in the selection of the best QTC-variant (QTC_B, QTC_C and their different levels and dimensions) for each of the sports. Furthermore, strongly linked to the previous challenge, is the decision regarding the different thresholds that can be used for the assignment of the QTC-characters. Finally, when analysing sports, the question arises on whether two identical movements patterns, where in the second occurrence players changed roles with respect to the first occurrence of the movement pattern, should be detected as identical. Regardless of the preference in this respect, this challenge can be addressed by including permutations of the QTC-representations into the distance calculation between the different QTC-matrix sequences.

Most of the challenges mentioned in this section can be seen as implementation tasks for the prototype, by the inclusion of parameters in the Python program (see Chapter 5). Consequently, by assigning different values to the parameters, the QTC method can be finetuned and calibrated when applying it for dynamic sports analytics. This can be done differently for each specific sport. They were programmed and will be discussed more in detail by means of an easy sports (three-cushion billiards) example in the first sections of Chapter 3. The method for the comparison of QTC-matrix sequences of different length, however, requires explanation that goes beyond such a simple introduction. Accordingly, this method (and its literature background) will be discussed more in detail in the last sections of Chapter 3. The reason for this is that it is a novel method that was introduced specifically for this research and published as a (general) research paper. Its applications, however, go beyond the use for sports analytics or for QTC analysis.

2.3.2 Static sports analysis

Static sports analysis entails the analysis of (average) positions at specific timestamps in sports and, in this research, makes use of a QTC-variant based on QTC_S that describes shapes (QTC_S). When applying this method to sports analytics of team sports, it can be seen as the analysis of team tactics, as tactics are mostly defined by positions of players in team sports (Rein & Memmert, 2016). The application of this variant for the sports analytics method requires different interpretation of QTC_S. In the original approach (Van de Weghe, 2004; Van de Weghe *et al.*, 2005), it is used to describe shapes (polylines and polygons) that, theoretically consist of an infinite number of points. The selection of the points defining such a shape is therefore a crucial part of the original method, before the actual description of these points by means of QTC_S. For the static sports analytics method, however, the points that will be described by QTC_S are given in advance (e.g. football or basketball players on a field who's positions at a specific timestamp need to be described). As such, the first step of the original approach can be neglected for the static sports analytics method. Furthermore, this implies that, in comparison to the points derived from polylines and polygons, the points of the static sports analysis do not have an explicit order. It is therefore impossible to construct a single defining shape based on the given points. For this reason, QTC_S for static sports analytics is named as 'QTC Static' rather than 'QTC Shape'. Following this line of reasoning, QTC_S therefore describes (sports) point formations rather than shapes (polylines and polygons). Consequently, whereas in the original approach QTC_S is used to describe the vectors between successive points (originating from the shapes), for the sports analytics method the relations between all of the possible vectors between all of the given points will be described. This difference results in a more precise description of the points but has some important downsides, that will be discussed to a further extent in Chapter 4, by means of a basic sports example. Furthermore, as the objects (e.g. football players) that are described with the QTC_S are moving though time, they are referred to as MPOs (Moving Point Objects) rather than POs, which was used as naming in the original approach.

Besides this, very similar to the dynamic sports analysis, some challenges arise when applying QTC_S for static sports analysis. A first challenge is the development of a suitable method that calculates distances between the QTC_S-descriptions of different point formations. In the original approach, distance was calculated by counting the relative number of different QTC-characters of the two QTC_S-description. It could be investigated whether a more detailed method is required to correctly conduct static sports analytics. The second challenge is the differentiation between sports point formations occurring at different absolute locations. Since QTC_S describes relative relations between the vectors between the points making up the sports point formation, two identical formations occurring at different locations, e.g. on the defensive or offensive half of a football field, are expected to have identical QTC_S-descriptions (and consequently a distance value of zero between them). Similar to the proposed adaptation for dynamic sports analytics (see Section 2.3.1), the idea is to include static points that can differentiate between these two point formations, into their respective QTC-descriptions and thereby increasing the resulting distance value between two such sports point formations. A third challenge is the decision regarding the different thresholds that can be used for the assignment of the QTC-characters that make up the QTC_S-description of a sports point formation. Finally, when analysing sports, the question arises on how to take players switching roles (and thus positions) into account and whether this is desirable when comparing two sports point formations. This issue can be addressed by taking permutations of the sport point formations into account when calculating the distance between them.

REFERENCES

Allen, J.F. (1981). An interval-based representation of temporal knowledge. In J.F. Allen (Ed), Proceedings of the 7th IJCAI (pp. 221-226). San Fransisco, CA, U.S.A.: Morgan Kaufmann.

Bera, R., Claramunt, C. (2003). Topology-base proximities in spatial systems. *Journal of Geographical Systems*, 5(4), 353-379.

Billen, R., Zlatanova, S. (2003). 3D Spatial relationships model: a useful concept for 3D cadastre? *Computers, Environment and Urban Systems*, 27, 411-425.

Bogaert, P., Van de Weghe, N., Cohn, A.G., Witlox, F., De Maeyer, P. (2007). The Qualitative Trajectory Calculus on Networks. In T. Barkowsky, M. Knauff, G. Ligozat, D.R. Montello (Eds), Spatial Cognition V Reasoning, Action, Interaction. Spatial Cognition 2006. Lecture Notes in Computer Science, vol 4387. Berlin, Heidelberg, Germany: Springer

Cohn A.G., Bennett, B., Gooday, J., Gotts, N. (1997). Qualitative Spatial Representation and Reasoning with the Region Connection Calculus. *GeoInformatica,* 1(3), 1-44.

Cohn, A.G. (2005). Qualitative spatial representation and reasoning techniques. *Lecture Notes in Computer Science,* 1303(1), 1-30.

Delafontaine, M., Chavoshi, S.H., Cohn, A., Van de Weghe, N. (2012). A Qualitative Trajectory Calculus to reason about moving objects. In M.S. Hazarika (Ed), Qualitative Spatio-Temporal Representation and Reasoning: Trends and Future Directions (pp. 148-168). Hershey, PA, U.S.A.: IGI Global.

Dijkmeijer, J., de Hoop, S. (1996). Topologic relations between fuzzy area objects. In M.J. Kraak, M. Molenaar (Eds.), Advances in GIS Researches II, Proceedings of the 7th Spatial data handling (pp. 377-394). London, U.K.: Taylor & Francis.

Dodge, S., Laube, P., Weibel, R. (2012). Movement similarity assessment using symbolic representation of trajectories. *International Journal of Geographical Information Science,* 26(9), 1563-1588.

Egenhofer, M.J. (1989). A formal definition of binary topological relationships. *Lecture Notes in Computer Science*, 367, 457-472.

Egenhofer, M.J., Herring, J.R. (1991). Categorizing binary topologic relations between regions, lines and points in geographic databases. Tech. Report, Dept. of Surveying Engineering, University of Maine, Orono.

Egenhofer, M.J., Shariff, A.R. (1998). Metric details for natural-language spatial relations. *ACM Transactions on Information Systems,* 16(4), 295-321.

Egenhofer, M.J. (2010). The Family of Conceptual Neighborhood Graphs for Region-Region Relations. *Geographic Information Science,* 6292, 42-55.

Freksa, C. (1992). Using orientation information for qualitative spatial reasoning. *Lecture Notes in Computer Science*, 639, 162-178.

Freksa, C., Zimmermann, K. (1992). On the utilization of spatial structures for cognitive plausible and efficient reasoning. In Proceedings of the 1992 IEEE International Conference on Systems Man and Cybernetics (pp.261-266). Chicago, IL, U.S.A.

Galton, A. (1990). A critical examination of Allen's theory of action and time. *Artificial Intelligence*, 42(2), 159-188.

Galton, A. (2000). Qualitative spatial change. Oxford, U.K.: Oxford University Press

Guan, L.J., Duckham, M. (2011). Decentralized reasoning about gradual changes of topological relationships between continuously evolving regions. In M. Egenhofer, N. Giudice, R. Moratz, M. Worboys (Eds.), Spatial information theory, Lecture notes in Computer Science. Berlin, Heidelberg, Germany: Springer.

Hazarika, S.M., Cohn, A.G. (2011). Qualitative spatio-temporal continuity. In D.R. Montello (Ed), Proceedings of the 5th International Conference on Spatial Information Theory (COSIT) (pp. 92-107). Morro Bay, C.L., U.S.A.: Springer.

Ibrahim, Z.M., Tawfik, A.Y. (2007). An abstract theory and ontology of motion based on the regions connection calculus. In I. Miguel, W. Ruml (Eds.), Abstraction, Reformulation, and Approximation (pp. 230-242). Whistler, Canada: Springer.

Kuipers, B. (1986). Qualitative simulation. *Artificial Intelligence*, 29, 289-388.

Kulpa, Z. (1997). Diagrammetric representation of interval space in proving theorems about interval relations. *Reliable Computing*, 3(3), 209-217.

Kulpa, Z. (2006). A diagrammetric approach to investigate interval relations. *Journal of Visual Languages and Computing*, 17(5), 466-502.

Kurata, Y. Egenhofer, M.J. (2009). Interpretation of behaviours from a viewpoint of topology. In B. Gottfriend, H. Aghajan (Eds.), Behaviour Monitoring and Interpretation-Ambient Intelligence and Smart Environments (pp. 75-97). Amsterdam: IOS Press.

Laube, P., Van Kreveld, M., Imfeld, S. (2004). Finding REMO – Detecting relative motion patterns in geospatial lifelines. In P.D. Fisher (Ed.), Developments in spatial data handling, Proceedings of the 11th international symposium on spatial data handling (pp. 201-214). Berlin, Heidelberg, Germany: Springer

Long, J.A., Nelson, T.A. (2013). Measuring dynamic interactions in movement data. *Transactions in GIS*, 17, 62-77.

Mavridis, N., Bellotto, N., Iliopoulos, K., Van de Weghe, N. (2015). QTC_{3D}: extending the qualitative trajectory calculus to three dimensions. *Information Sciences*, 322(1), 20-30.

Moreira, J., Ribeiro, C., Saglio, J.M. (1999). Representation and manipulation of moving points: an extended data model for location estimation. *Cartography and Geographic Information Systems*, 26(2), 109-123.

Morgulev, E., Azar, O.H., Lidor, R. (2017). Sports analytics and the big-data era. *International Journal of Data Science and Analytics,* 1, 1-10.

Muller, P. (2002). Topological spatio-temporal reasoning and representation*. Computational Intelligence*, 18(3), 420-450.

Randell, D., Cui, Z., Cohn, A.G. (1992). A Spatial Logic Based on Regions and Connection. In B. Nebel, W. Swartout, C. Rich (Eds.), Proceedings of the 3rd International Conference on Knowledge Representation and Reasoning (pp. 165-176). San Mateo, CA, U.S.A: Morgan Kaufman.

Rein, R., Memmert, D. (2016). Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. *Springerplus,* 5(1), 1410.

Van de Weghe, N. (2004). *Representing and Reasoning about Moving Objects: A Qualitative Approach*. Ph.D. dissertation, Ghent University.

Van de Weghe, N., Cohn, A.G., De Maeyer, P., Witlox, F. (2005). Representing moving objects in computer-based expert systems: the overtake event example*. Expert Systems with Applications*, 29(4), 977-983.

Van de Weghe, N., Cohn, A.G., De Tré, G., De Maeyer, P. (2006). A Qualitative Trajectory Calculus as a basis for representing moving objects in Geographical Information Systems. *Control and Cybernetics,* 35(1), 97-120.

Van de Weghe, N., De Maeyer, P. (2005). Conceptual neighbourhood diagrams for representing moving objects. *Lecture Notes in Computer Science,* 3770, 228-238.

Van de Weghe, N., De Tré, G., Kuijpers, B., De Maeyer, P. (2005). The double-cross and the generalization concept as a basis for representing and comparing shapes of polylines. In R. Meersman, Z. Tari, P. Herrero (Eds.), On the move to meaningful internet systems. Lecture Notes in Computer Science (pp.1087). Berlin, Heidelberg, Germany: Springer.

Von Landesberger, T., Brodkorb, F., Roskosh, P., Adrienko, N., Adrienko, G., Kerren A. (2016). Mobilitygraphs: Visual analysis of mass mobility dynamics via spatio-temporal graphs and clustering. *IEEE Transactions on Visualisation and Computer Graphics*, 22, 11-20.

Williams, M.J., Musolesi, M. (2016). Spatio-temporal networks: Reachability, centrality and robustness. *Royal Society Open Science*, 3(6), 160-196.

Worboys, M.F. (1994). A Unified model for spatial and temporal information. *The Computer Journal,* 37(1), 26-34.

Zhang, P.D., Beernaerts, J., Zhang, L., Van de Weghe, N. (2016). Visual exploration of match performance based on football movement data using the continuous triangular model. *Applied Geograpphy*, 76, 1-13.

Zhang, P.D., Beernaerts, J., Van de Weghe, N. (2018). A hybrid approach combining the multi-temporal scale spatio-temporal network with the continuous triangular model for exploring dynamic reactions in movement data: a case study of football. *ISPRS International Journal of Geo-Information*, 7(1), 1-18.

3.

QTC FOR DYNAMIC SPORTS ANALYTICS

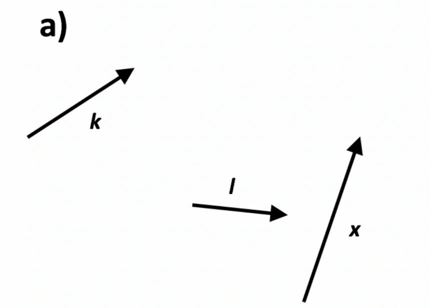
Parts of this chapter were modified from: BEERNAERTS, J., DE BEVER, E., LENOIR, M., DE BAETS , B., VAN DE WEGHE, N. 2019. A method based on the Levenshtein distance metric for the comparison of multiple movement patterns described by matrix sequences of different length. *Expert Systems with Applications,* 115, 373-385.

In this Chapter, the proposed methodology for dynamic sports analytics is introduced. This method supports the analysis of team behaviour with QTC. First, the transition from QTC-relations describing movement between two MPOs during one time interval, to the representation of movement of multiple MPOs during multiple time intervals is discussed. Several other concepts need to be included into the prototype of the program (see Section 2.3.1), to be able to use QTC for dynamic sports analytics. Most of these concepts *(i.e.* static points, thresholds for QTC-characters, different QTC-variants or permutations), however, do not consist of big methodological changes or additions to the QTC-methodology, but rather of code implementation. These concepts are therefore introduced at the beginning of this chapter. Rather than giving a theoretical overview, we illustrate these adaptations and changes by means of three-cushion billiards examples. The reason for this is that the goal is to give examples by means of a relatively simple sport before switching to more complex sports in the use cases of Chapters 6, 7 and 8 of Part II of this dissertation. With three-cushion billiards, for example, there are only three balls (thus three MPOs), of which the movement is only defined by an initial shot and the laws of nature (e.g. gravity, friction). Therefore, once the shot has been placed, there are no more human factors involved such as tactics, which results in movements with lower complexity that are easier to predict. The applications (and finetuning) of the different additions, are discussed more in detail for specific sports discussed in the use cases (Chapters 6,7 and 8).

Thereafter, distance calculation between different movement patterns is discussed, based on basic sequence alignment methods. These basic methods, however, cannot be used for distance calculation between (QTC-)matrix sequences of different length. Since the aim is to enable the comparison of sport events with different temporal lengths, one major addition to the QTC methodology is required. For this, a new method is proposed. We include a more thorough example where the method is used for the dynamic analysis of a series of three-cushion-billiards opening shots. The chapter ends with a critical reflection on the different possible applications of this novel method, within sports and other domains such as geography. Unfortunately, within the research presented in this dissertation, no solution for the comparison of sports events containing different numbers of players was found and this should be considered as future work.

3.1 Describing multiple MPOs and time intervals

As described in Chapter 2, QTC_B and QTC_C only take two MPOs into account during a specific time interval. When considering the movement of multiple MPOs (Figure 3-1a) with QTC_{C21}, a matrix can be used to collect all pairwise comparisons between the different objects (Figure 3-1b). An element (*i*,*j*) of this matrix represents the QTC_{C21}-relation between MPOs *i* and *j* (Van de Weghe *et al.*, 2005). The movement interaction between vectors *k* and *l* during the time interval *T*(1-2) can be described by a '-+-+'-relation in QTC_{C21}. The first character (-) indicates that vector *k* is moving towards the starting point of vector *l*, the second character (+) indicates that vector *l* is moving away from the starting point of vector *k*. The third and fourth characters indicate the directions of the vectors compared to the reference line connecting the starting points of both vectors, as described in Chapter 2.



b)

T(1-2)	k	l	x
k		- + - +	- - - +
l			- - - +
x			

Fig. 3-1. Describing multiple MPOs with QTC_{C21} by means of a QTC_{C21}-matrix.

When the movement of multiple objects is described over longer time, a sequence of matrices can be used to store the QTC_{C21}-relations for all the different time intervals. Real-world sports movement interactions can thus be described by means of matrix sequences, where the number of matrices in the sequence is dependent on the temporal duration of the event. This is why, when comparing a series of similar events such as the overtake example (Van de Weghe *et al*., 2005), small differences can occur in the length of the QTC_{C21}-representations. These can be caused by differences inherent to the spatial movement patterns of that event, but also because identical trajectories were performed at different speeds when they are described in an equal-interval based manner. In sports this occurs often, as similar movement patterns, e.g. an attack in football, seldomly have exact the same temporal lengths.

The movements of a set MPOs during a specific well-defined time interval of interest will be referred to as 'a fragment' from here on. As such, each fragment can be described by a QTC-matrix sequence, where the number of players determines the dimension of the individual QTC-matrices that make up the sequence, while the sequence length is defined by the temporal length and resolution of the fragment.

3.2 Introduction of the three-cushion billiards example

Three cushion billiards is a variant of the more popular carom billiards. The game is played with three balls, at start positioned as displayed in Figure 3-2. The goal of the game is for the player to play the cue (white) ball and to hit the two object balls (red and yellow), while hitting three of the sides with the cue ball before touching the last object ball. According to the success of the shot, points are given to the player. After the first shot, the game continues with the three balls positioned as they were at the end of the first shot (Cohen, 2002). Three cushion billiards was chosen to serve as an example for the methodology because of its simplicity. First of all, there are only three balls (thus three MPOs), of which the movement is only defined by an initial shot and the laws of nature (e.g. gravity, friction). Therefore, once the shot has been placed, there are no more human factors involved e.g. tactics, physical condition, physical capabilities or psychological effects. Consequently, three-cushion billiards is characterised by movements that have lower complexity than most other (team) sports. Similarly, their movement patterns of the balls are easier to predict.

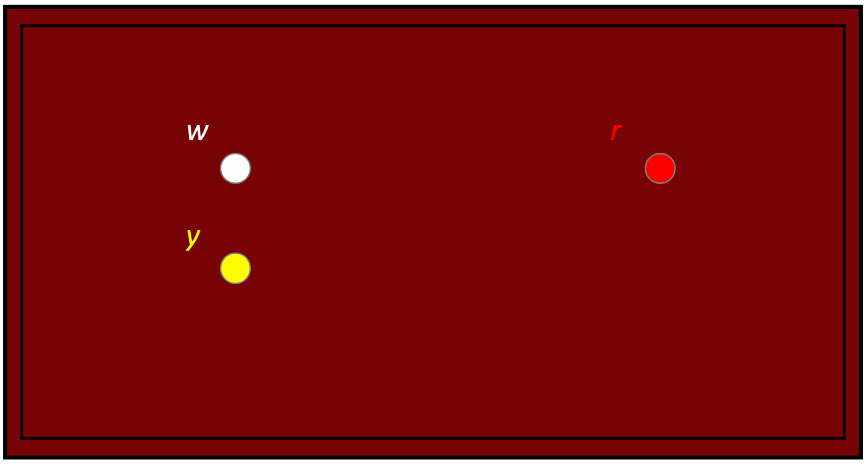


Fig. 3-2. The starting formation of three-cushion billiards with three billiard balls *w*, *y* and *r*.

For the examples used in this chapter, only the opening shot of the three-cushion billiards is considered. A series of 55 shots were generated by an online tool (http://www.casualarena.com/games-french.php), by applying the most popular opening shot technique used by casual players (Byrne, 1998). The (*x*,*y*,*t*)-coordinates of the centroids of the three balls were logged with a temporal resolution of 20Hz. The 55 shots get named by the number of the shot (1 to 55), followed by a letter (S, R or L) according to the type of the shot (see Figure 3-3). Fourteen of the 55 opening shots are from the type 'success' (label 'S'), meaning that the goal of hitting the two object balls and three sides was achieved. The other 41 shots are divided in two types, according to their characteristics and their trajectories. The 'pass on the right'-type (label 'R') contains shots where the white ball hit the red ball but passed on the right of the yellow ball. It contains a subgroup called 'yellow touched' of shots with similar properties, with the difference that the red ball hit the yellow one. Shots of the 'pass on the left'-type (label 'L') consist of the white ball first hitting the red one and thereafter passing on the left of the yellow ball.

Type	Group label	Number	Average temporal length (seconds)	Stdev	Average temporal length (timestamps)	Stdev
success	S	14	5.01	0.07	100.14	1.35
pass on the right (+ yellow touched)	R	23	5.00	0.09	100.04	1.82
pass on the left	L	18	5.18	0.13	103.50	2.64

Fig. 3-3. Overview of the three-cushion billiards dataset.

3.3 Using static points

When two identical movement patterns occur at different absolute locations, e.g. at different locations on the billiards table, their QTC-descriptions will be identical, as is shown for two fragments for QTC_{B21} in Figure 3-4. The reason is that only relative movements between the MPOs making up that movement pattern are taken into account. In this case, this is the relative movement between the white (w) and yellow (y) balls. For reasons of simplification, only these two balls were included in the example. To a billiard player, however, the two movements presented in fragment 1 and fragment 2 of Figure 3-4 will have an intrinsic difference, despite the fact that they are visually very similar. Because of a difference in absolute positioning (relative to the table), the context of the movements in both fragments is different, for example because their end positions on the table after the movement presented in both fragments influences the tactics that should be used in the next shot.

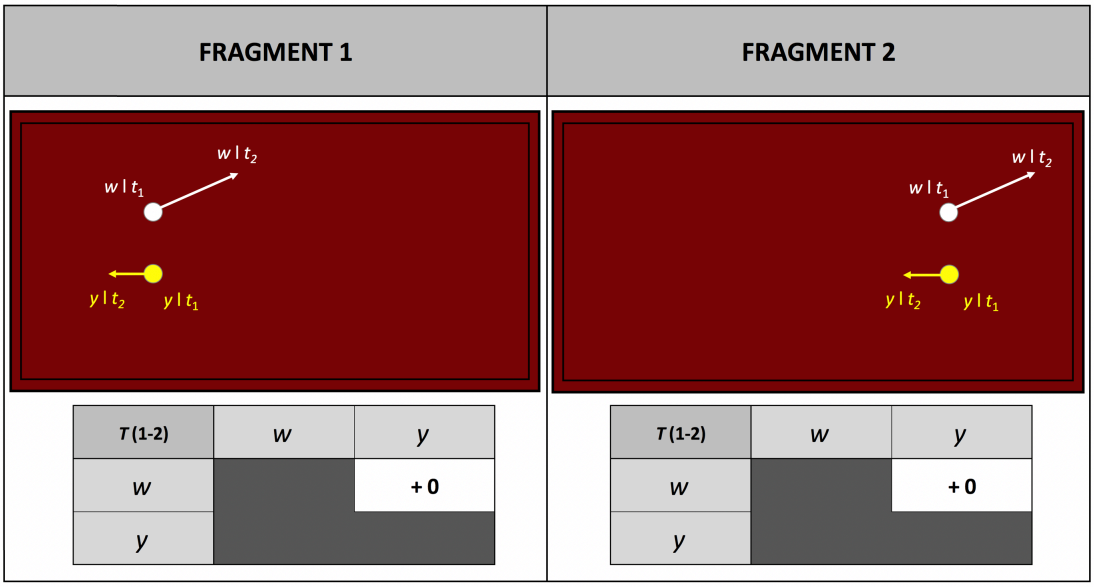
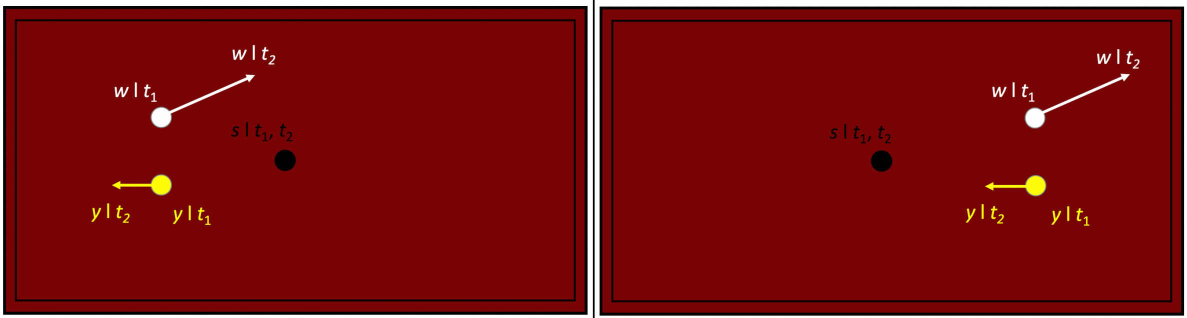


Fig. 3-4. Two billiard fragments containing the movement of two billiard balls *w* and *y* during one time interval *T* and their identical QTC_{B21}-representations.

To distinguish between two identical movement patterns occurring at different absolute locations, a static point can be included into the QTC-description. In Figure 3-5, a static point *s* is added to the two fragments that were previously shown in Figure 3-4. Along the visualisations, the QTC-descriptions of the two fragments, including the static point *s*, are shown. It can be seen that the QTC-descriptions are no longer identical, as the static point, which is placed at the middle of the billiards table, differentiates between movements that are going towards or away from the centre of the table. Consequently, movements happening on the left or on the right of the billiards table can be discerned by adding points on the corners of the billiards table.

FRAGMENT 1	FRAGMENT 2



Fragment 1:

T (1-2)	w	y	s
w		+ 0	- 0
y			+ 0
s			

Fragment 2:

T (1-2)	w	y	s
w		+ 0	+ 0
y			- 0
s			

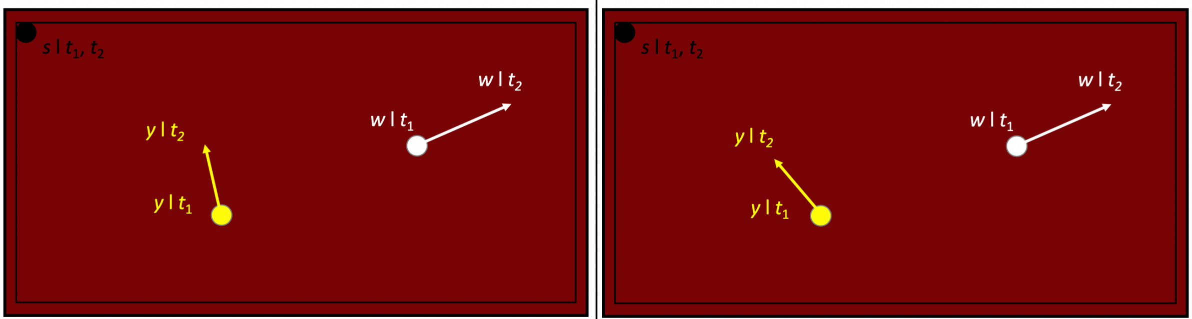
Fig. 3-5. Two billiard fragments containing the movement of two billiard balls *w* and *y* and one static point *s* during one time interval *T* and their non-identical QTC_{B21}-representations.

Similarly, (multiple) static points can be added to the QTC-descriptions of movement patterns of more complex sports, to take absolute locations into account. Moreover, static points can be used to enhance the description of the movement of limited numbers of MPOs. When describing the movement of a limited number of MPOs, there are only a limited number[7] of QTC-relations for the description of each MPO. Since the QTC-characters of a QTC-relation can only belong to the threefold domain D = {-, 0, +}, this can result in a QTC-description with a low level of detail. In certain cases, such as in Figure 3-6, small differences in such a movement pattern can result in largely different QTC-matrices. By adding one or multiple static points to the QTC-description (and thus adding more QTC-relations to the description of the movement), the relative weight of the QTC-relations between the MPOs in the total QTC-description can be reduced. This is important when calculating an appropriate distance between two QTC-representations (Section 3.7). In Figure 3-6, two fragments and their QTC_{B21}-descriptions are shown. The only difference between fragment 1 and fragment 2 is a small difference in the movement direction of MPO *w*. When the static point *s*

[7] When describing the movement of *n* MPOs with QTC_{B21}, each MPO is described by *n*-1 QTC_{B21}-relations

would not be included in the QTC$_{B21}$-description of the fragments, this small direction difference would result in a 50% difference between both QTC$_{B21}$-descriptions ('+ -' for fragment 1 and '+ +' for fragment 2). For both fragments, however, it can be seen that the QTC$_{B21}$-relations of the MPOs w and y with respect to the static point *s* are identical. As such, static point can be used to finetune the QTC-description of movement patterns consisting of a limited number of MPOs.

FRAGMENT 1	FRAGMENT 2



Fragment 1:

T (1-2)	*w*	*y*	*s*
w		+ -	+ 0
y			- 0
s			

Fragment 2:

T (1-2)	*w*	*y*	*s*
w		+ +	+ 0
y			- 0
s			

Fig. 3-6. Two billiard fragments containing the movement of two billiard balls *w* and *y* and one static point *s* during one time interval *T* and their QTC$_{B21}$-representations, where only the movement of *y* differs slightly in both fragments.

In Figure 3-6 it can be seen that the use of the static point *s* results in invariant '0'-characters in the QTC-relations between *s* and each of the MPOs. It is important to note that, because of the strict definition that a static point does not move through time, these characters are invariant. There will therefore always be a '0'-character in the QTC-relation of a static point with respect to other MPOs or other static points (in the case of the usage of multiple static points). When calculating distances between two QTC-descriptions, these invariant characters should therefore not be taken into account (see Section 3.8.2), to obtain accurate distances. If not, the distance that is calculated between two fragments would lower gradually while more static points are added to these two fragments, because of the proportional increase in identical characters in both QTC-representations. Since only one distance is calculated between both fragments, including the invariant characters in the distance calculation could reduce the impact of the QTC-characters that correspond with significant real-world movement differences, through over-averaging. Moreover, when using multiple static points, the QTC-relations describing the relative movement between two static points should be omitted from this distance calculation, since all of the QTC-characters of that cell of the QTC-matrix are invariant and will be '0' by definition.

Another remark is that by the introduction of static points that are placed close to or inside the bounding box of the movements, such as in the examples of this section, the movement patterns remain rotation invariant. This means that a rotation of the movements with respect to the static point (can) result(s) in an identical QTC-description. For static points placed closer to the centre of the bounding box, the maximum angle of the rotation invariance (expressed between 1° and 360°) is larger than for points placed farther away from this centre. To enforce rotation variance, a static point placed at infinity can be included.

3.4 Using different QTC-variants

A second challenge is the selection of the most suitable QTC-variant. In general, this is a decision that needs to be supported by testing the different variants for each different application through the description and comparison of the fragments of a sample dataset by means of QTC. Subsequently, the results (distances between the fragments) of the different variants can be compared with the ground-truth, consisting of the actual distances between the fragments. The most suitable QTC-variant can be selected by choosing the results that best correspond with the ground truth. In such a test, QTC_C rather than QTC_B turned out to be more suitable to be used for three-cushion billiards analysis. However, since such tests often require a lot of time and a ground-truth might be lacking, some guidelines can be used to select the most suitable QTC-variant.

First of all, it is advisable to use QTC_C when dealing with limited numbers of MPOs (and optionally, static points) or when movement directions in particular are very important. The more MPOs and static points are included in the fragment, the more detailed the QTC-description will define the movements occurring in that specific fragment. This is because, when taking one MPO as reference, each of the n-1 other MPOs and static points puts a constraint on the possible movement direction of the reference MPO by means of a QTC-character. The different movement constraints are combined, resulting in a strictly defined movement direction for each of the MPOs. In Figure 3-7, the incrementally decreasing possible movement direction of one billiard ball y for one time interval T is shown by stepwise addition of two other billiard balls (w and r) and two static points (s_1 and s_2) and their respective QTC_{B21}-relation with respect to y. Note, however, that not every additional MPO results in a further restriction of the possible movement direction, especially when the possible movement direction is already rather limited, which is shown by the addition of s_1 in Figure 3-7. Generally, it can be advised to use the more detailed QTC_C rather than the simpler QTC_B when describing a fragment containing limited numbers of MPOs, to ensure a detailed description of the movements occurring in that fragment. This is the most important reason that QTC_C turned out to be most suitable to analyse three-cushion billiards. Furthermore, for applications where a fine description of movement directions is extremely important, e.g. in Formula 1 where success is often defined by the capabilities of a driver to use the slip stream of other cars or perform good overtakes. On the downside, QTC_C requires more processing power and storage (see Chapter 5) and might result in over-averaging when calculating distances between fragments if the difference in movement is not expressed by the extra characters of the more detailed variant (see Section 3.3). For these reasons, one might opt for QTC_B when describing movements of a large number of MPOs or for description of large datasets.

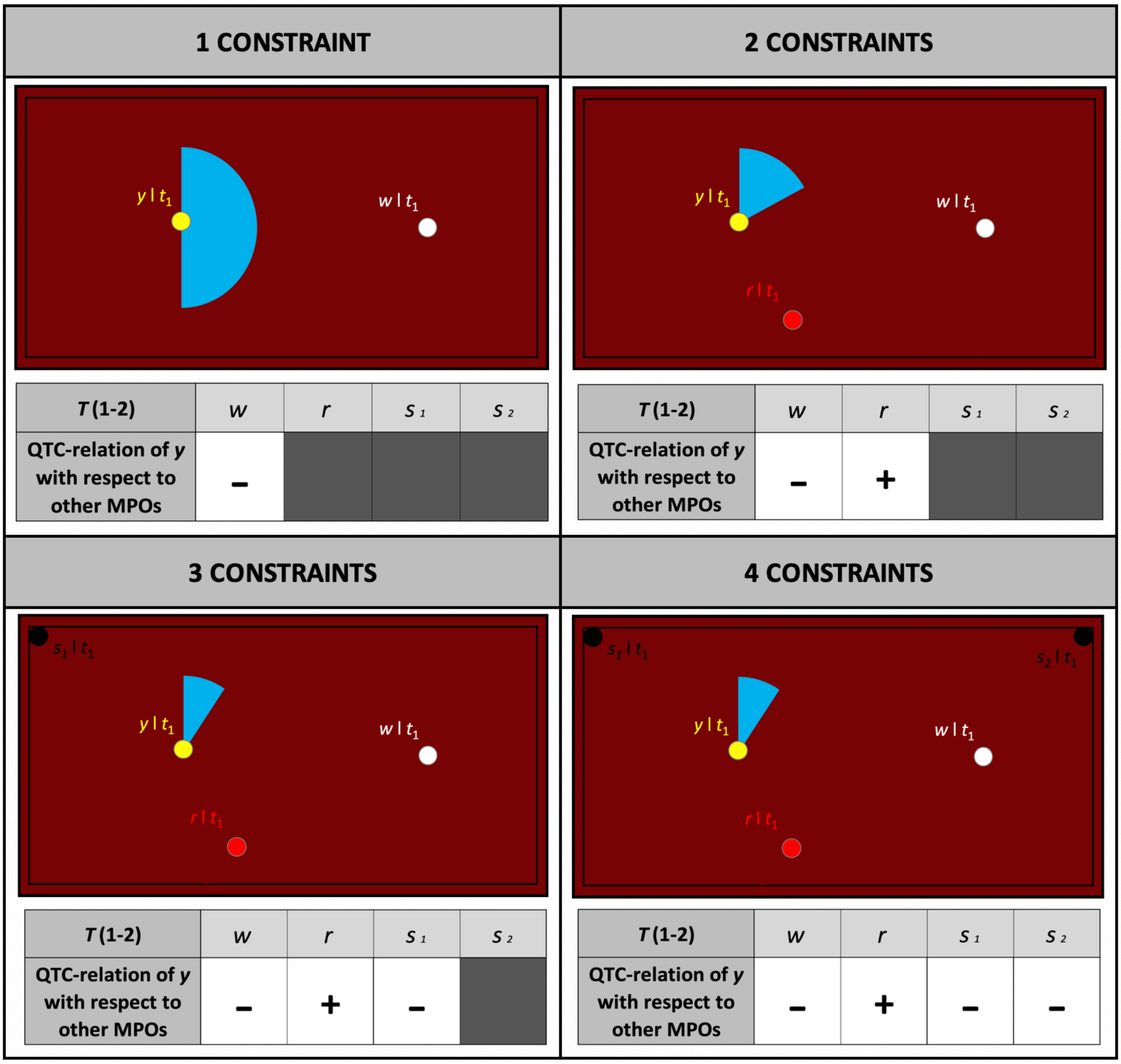


Fig. 3-7. The impact of combining multiple QTC-movement constraints of two billiard balls *w* and *r* and two static points s_1 and s_2 on the possible movement direction (shown in light blue) of a yellow billiard ball *y* during one time interval *T*.

A second guideline is that the dimension of the QTC-variant should be chosen by a domain expert, as this defines the movements that will be subject of the study. For most sports, however, it is straightforward to opt for 2-dimensional analysis (e.g. billiards, football, cycling, etc.) whereas in other sports (e.g. swimming, athletic jumping, etc.) a 3-dimensional approach might be advisable. In the 2-dimensional case it is essential to choose the most suitable plane (*x*,*y*; *x*,*z*; *y*,*z*), since sports data is increasingly logged for three dimensions. Regardless of the choice of dimensionality, it might be advisable to start by analysing more complex movements in 2-D before switching to 3-D, to first gain insights in the capabilities of the considered QTC-variant for the specific application.

3.5 Using angular tolerances for the assignment of the QTC-characters

Another challenge of applying the QTC methodology on real-world situations is the unequal relative occurrences of specific QTC-relations in the real world. For example, the strict constraint for the 'o'-character for level one QTC_B makes its occurrence very unlikely in sports situations described by QTC. Although meaningful, the 'o'-character is unlikely to occur in a level one QTC_B-relation between MPOs because of a number of reasons. The first reason is, as mentioned before, the strict constraint for the assignment of the character, which can be seen in Figure 2-3 of Chapter 2. The second reason is that the limited (spatial) accuracy of the tracking methods makes it unlikely that the movement of an MPO during one time interval T is directed exactly at another MPO. Third, because of the limited temporal resolution that can be achieved for logging sports data (see Section 2.3.1), the intermediate 'o'-character is not guaranteed to occur for transitions between '-' and '+'-characters. Similarly, the 'o'-character is unlikely to occur for the four characters of the level one QTC_C in sports situations. The first two characters are identical to the QTC_B-relation, the latter two describe the movement of the MPOs according two reference lines (see Section 2.2.2) and are as such prone to the same issues as described above.

For these reasons, an angular tolerance α for the assignment of the 'o'-characters in level one QTC_B and level one QTC_C could be allowed, as shown in Figure 3-8 for the 2-dimensional case. This is the allowed deviation from both sides of the reference line for assignment of the 'o'-character. Such an angular tolerance could ensure that the 'o'-character is assigned more often, or to make sure that the movement of an MPO with a movement direction with a minimal deviation from the 'o'-line gets the same QTC-description as a movement with a direction that exactly corresponds with this line, assuming that they correspond to very similar real-world (sports) situations. Note that the angular tolerance α does not need to be identical for the level one QTC_B as for the two last characters of the level one QTC_C. In Figure 3-8, however, α is set at 15° for both cases.

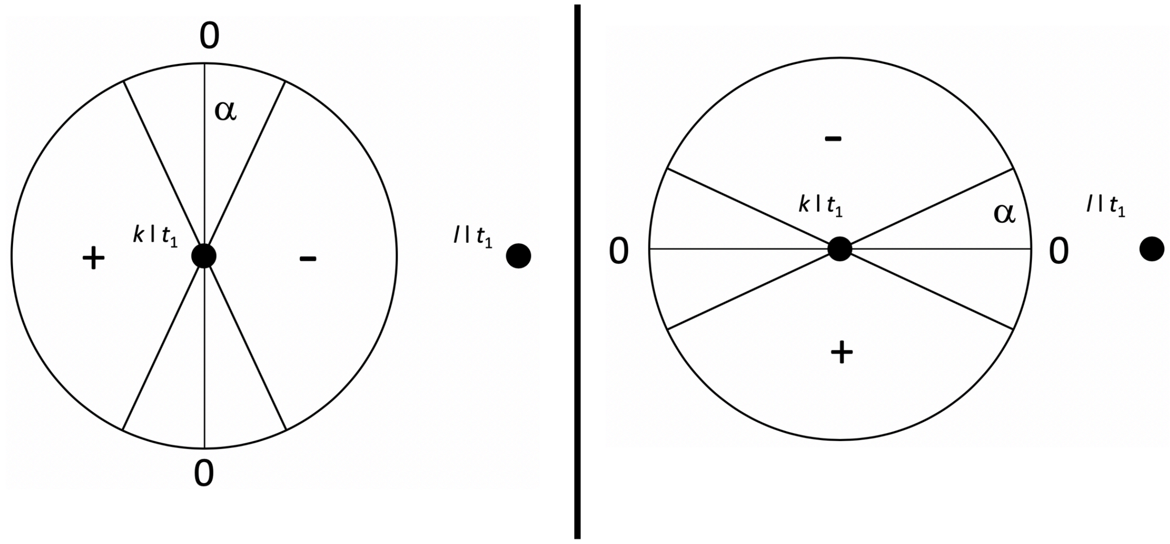
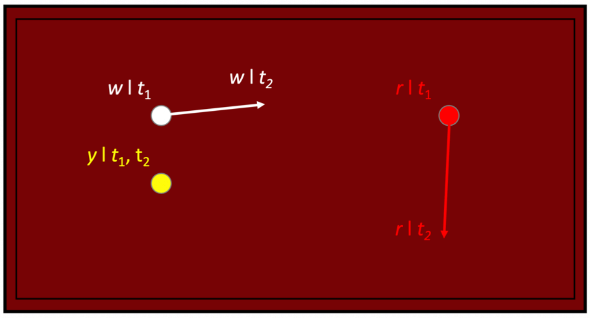


Fig. 3-8. The concept of angular tolerance α for the 'o'-character for level one QTC_B for *k* with respect to *l* (left) and for the 'o'-character of the two last characters of level one QTC_C, with α =15°.

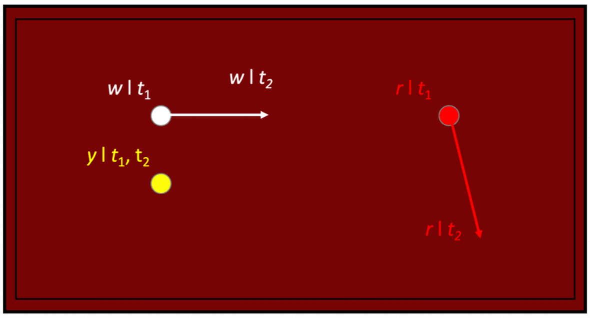
Including such an angular tolerance, however, might cause some unexpected and often unwanted consequences. First, it might result in a lower number of assignments of '-' and '+'-characters, and, consequently, lower distances between different fragments. Secondly, since the angular tolerance is defined as a fixed angle starting from the reference MPO, its impact on the resulting QTC-character is influenced by the distance of the other MPOs to that reference MPO. The closer another MPO is, the smaller the allowed deviation in Euclidean distance of the end of the movement vector of the reference MPO. Consequently, the speed and the temporal resolution of the dataset also influence the impact of the tolerance angle on the resulting QTC-character. The last and most important problem, however, is the possibility of unwanted changes to QTC-relations between other MPOs that were not intended when including the angular tolerance. As shown in Figure 3-9, one might want to see the real-world similarities of the movements in fragment 1 and fragment 2 reflected in the respective QTC_{B21}-descriptions. Although the two fragments show small differences in the movement of the billiard balls *w* and *r* during one time interval *T*, one could find these differences neglectable and would want to obtain a distance of zero between the QTC_{B21}-representations of both fragments. However, it can be seen from the top part of the figure that two of the cells in the QTC_{B21}-matrices describing the QTC_{B21}-relations between *w* and *y* on the one hand and *w* and *r* on the other hand, contain different QTC_{B21}-characters. Because of this, the calculated distance will be higher than zero. To try and solve this issue, an angular tolerance (e.g. α =10°) could be taken into account while calculating the QTC_{B21}-representations. Although this results in equal QTC_{B21}-relations between *w* and *y* on the one hand and *w* and *r* on the other hand, it also results in an unexpected difference in the QTC_{B21}-relation between *y* and *r*. The angular tolerance will therefore not result in a distance equal to zero.

The example presented in Figure 3-9 illustrates that it is not always straightforward to interpret all the consequences of including an angular tolerance for the assignment of QTC-characters. While it might be suitable to increase the performance of QTC to detect similarities in certain situations, it is not advisable for the comparison of movements of a large number of MPOs or a large number of fragments, as the interpretation of the side-effects might not be straightforward in such cases. Furthermore, a clear goal of including the angular tolerance is required. In traffic, for example, this could be the detection of vehicles driving in lanes, *i.e.* small movements within the lane could be ignored by including an angular tolerance, whereas lane changes should result in different QTC-relations with respect to other vehicles. For sports analytics, the method could be used in a similar manner for Formula 1, to better detect the difference between the overtake of one car by another on the one hand and small movements of the front car on the other hand, which are done to make it difficult for the car behind him to take advantage of his slipstream. Hence, a parameter to define an angular tolerance for the 'o'-character for the calculation of the QTC_B- and QTC_C-matrices was included in the Python program (see Chapter 5) but was very rarely used for the case studies conducted in this dissertation.

FRAGMENT 1



FRAGMENT 2



WITHOUT TOLERANCE ANGLE

Fragment 1:

T (1-2)	*w*	*y*	*r*
w		+ 0	- -
y			0 -
r			

Fragment 2:

T (1-2)	*w*	*y*	*r*
w		0 0	- +
y			0 -
r			

WITH TOLERANCE ANGLE

Fragment 1:

T (1-2)	*w*	*y*	*r*
w		0 0	- 0
y			0 -
r			

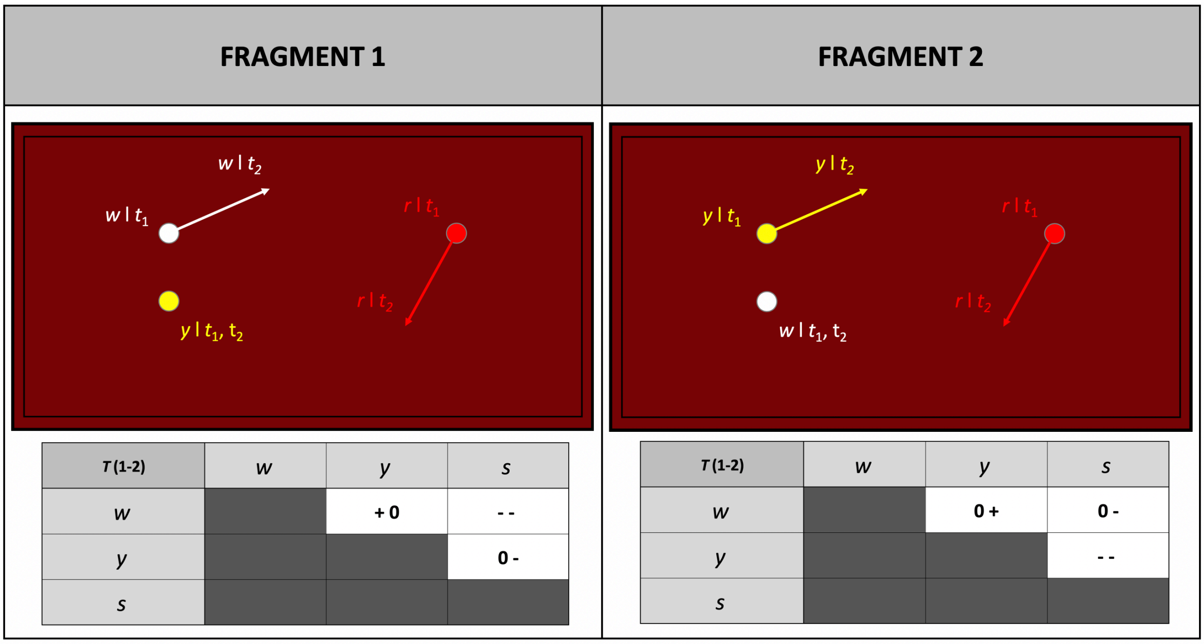
Fragment 2:

T (1-2)	*w*	*y*	*r*
w		0 0	- 0
y			0 0
r			

Fig. 3-9. The impact of taking an angular tolerance (α =10°) into account for the calculation of the QTC_{B21}-representations of two highly similar fragments including the movements of three billiard balls *w*, *y* and *r* during one time interval *T*.

3.6 Using permutations

A fourth challenge lies in the decision on the importance of the identity of the different MPOs of a fragment. In layman's terms this can be described by the following question: "Is it important who (e.g. which player, ball, etc.) performs which part of the movement that makes up the movement pattern of a fragment?". More precisely, should two fragments be considered as identical when all of the movements in them are identical but if two or more MPOs take up each other's place? An example showing two such fragments containing identical movements performed by different MPOs is shown in Figure 3-10. Both fragments contain the movement and QTC_{B21}-representations of three billiard balls *w*, *y* and *r* during one time interval *T*. In the second fragment, however, the roles of *w* and *y* are switched, while all other movements stay identical. Consequently, this difference can be seen in the respective QTC_{B21}-matrices, with different relations between each of the MPOs shown at the bottom of Figure 3-10. Despite the visual correspondence of the movements in both of the fragments, the comparison of the QTC-descriptions would therefore result in a high distance (see distance calculation in Section 3.7). In the prototype of the program this is solved by including a parameter that lets the user define whether these differences should result in a high distance or not (see Chapter 5). In the first case, the normal distance calculation is executed (see Section 3.8), for the latter, the distance calculation between the two fragments includes the permutations of the respective QTC-matrices.



Fragment 1:

T (1-2)	*w*	*y*	*s*
w		+ 0	- -
y			0 -
s			

Fragment 2:

T (1-2)	*w*	*y*	*s*
w		0 +	0 -
y			- -
s			

Fig. 3-10. The problem of permutations for dynamic sports analytics. Two billiard fragments containing the movement of three billiard balls *w*, *y* and *r* during one time interval *T* and their QTC_{B21}-representations.

3.7 Distance calculation between different movement patterns

For comparing different (sport) movement events, the distance is calculated between the qualitative representations of these events. Note that this distance is not the spatial distance but rather a measure of dissimilarity between the qualitative representations of those movement events. After obtaining these distances, a similarity analysis can be performed (Gentle, 2007). This similarity analysis usually results in a grouping of the most similar events and can be elucidated by means of visual representations such as a dendrogram in hierarchical clustering. In this section, we will present some of the most common (basic) techniques used for distance calculation between sequences. First of all, an introduction to basic alignment methods is given. These assume that the elements of the sequences are fixed, meaning that they are elementary building blocks of the sequence and cannot be further divided, and are finite in number. A simple and common example of such fixed building blocks of sequences are the letters of a word. In the remainder of this chapter these undividable building blocks of sequences are referred to as “sequence elements”. In case of the basic alignment methods, pairwise distance calculation can be performed based on a conceptual distance table. Secondly, an introduction of distance calculation between matrix sequences is provided. Where the former methods can calculate distances for sequences of different lengths, current approaches for matrix sequences do not have this ability. However, when describing similar sport events (*i.e.* fragments) by means of a qualitative calculus such as QTC, the qualitative representations, in this case sequences of matrices, do not necessarily have the same length, even if the user would classify them as similar based on a visual analysis of the movements.

3.7.1 Basic sequence alignment methods

Sequence Alignment Methods (SAMs) are methods that equate two or more sequences of elements by applying a set of qualified operations (Morrison, 2010). Such methods try to achieve optimal alignments by employing dynamic programming algorithms to minimise a distance metric (Wilson, 2008). There are two types of sequence alignment algorithms: global alignment methods force alignment to cover the entire length of the sequences, while local alignment methods base alignment on similar parts in the sequences, even if the entire sequences differ significantly (Schlich, 2003). Important to note is that basic sequence alignment methods enable alignment of sequences of different length, opposed to the so-called lock-step measures (Ranacher & Tzavella, 2017). The need for SAMs for calculating distances between sequences arose in biology, where sequences of DNA (or RNA, proteins, etc.) were subject of similarity analysis (Rosenberg, 2009). A breakthrough in the computation of sequence alignment occurred in 1966 when Levenshtein introduced a distance metric based on edit operations (Levenshtein, 1966). Numerous new methods for sequence alignment were proposed over time, taking the Levenshtein method as a starting point. As the novel method proposed in this chapter also takes the Levenshtein method as base, a thorough overview of the technique is provided here. A good understanding of the Levenshtein method is therefore crucial for the comprehension of the remainder of the chapter and the given examples. Levenshtein calculates a distance based on edit operations, commonly called the *edit distance,* which can be defined as the minimum cost of changing one string into another by applying a sequence of weighted edit operations usually including *identity*, *substitution*, *insertion* and *deletion*. As the latter two always occur together, they are usually called *indels.* As such, these basic sequence alignment methods are capable of aligning sequences of different length.

The suitable operations have specific costs that are chosen by the user. In terms of distance, the identity operations typically have no cost, while substitutions and *indels* are usually associated with positive costs. To display the editing process, the gap representation (Figure 3-11) is often used. This representation places strings one above the other, with a gap in the second string for every insertion corresponding with a deletion in the first string (Wilson, 2008). As can be seen in Figure 3-11, a minimum of three steps is needed to go from [BILLIARD] to [BILJART].

B	I	L	L	I	A	R	D
B	I	L	_	J	A	R	T
✔	✔	✔	-	✘	✔	✔	✘

✔ **Identity**
✘ **Substitution**
\- **Indel**

Fig. 3-11. Pairwise alignment of the English word 'BILLIARD' and its Dutch translation 'BILJART'. (after Delafontaine & Versichele *et al.*, 2012)

For longer examples, it is considerably more difficult to define the minimum number of steps and hence algorithms are needed to calculate the distance. One of the most popular algorithms for the computation of the edit distance between two strings, based on dynamic programming, is the algorithm of Needleman & Wunsch (1970). Despite the vast number of algorithms for edit distance calculation, it can be seen as a generally accepted standard (Waterman, 1984; Tönges *et al.*, 1996). The Needleman & Wunsch method breaks down the original sequences into a series of smaller elements. The idea is to construct the best alignment by using optimal alignments of the smaller subsequences, while guaranteeing that the optimal alignment solution will be found. Pairwise alignment is basically the process of equating a source and a target string using a set of accepted operations (Wilson, 2008). Following the Needleman & Wunsch method, the first step is to write the elements of the so-called source and target sequence in the margins of a comparison table of $n + 1$ rows and $m + 1$ columns (Figure 3-12), with n and m respectively equal to the number of characters in the source and target sequence. A cumulative cost is calculated beginning at the upper left cell (0,0), proceeding to the lower right cell ($m + 1$, $n + 1$). A diagonal movement represents either an identity or substitution operation. A horizontal movement in the table represents an insertion in the source string or a deletion in the target string. A vertical movement in the table represents a deletion in the source string or an insertion in the target string. The minimum edit distance for the alignment of the two strings can be found in the lower right cell ($m + 1$, $n + 1$) of the comparison table.

Having an m-character source string $A[1\ldots m]$ and an n-character target string $B[1\ldots n]$, every element $E(i,j)$ in the comparison table is the edit distance between the prefixes of length i and j, *i.e.* $A[1\ldots i]$ and $B[1\ldots j]$. The four possibilities (identity, insertion, deletion and substitution) are evaluated and the minimum value among the four operations is chosen. Note that the substitution cost is zero in the case of identical characters and the cost for insertion is equal to the cost of a deletion (written as *indel*).

$$E(0,0) = \ \ 0, E(i,j) = \min \begin{cases} E(i-1, j-1) & + \textit{substitution cost} \\ E(i, j-1) & + \textit{indel cost} \\ E(i-1, j) & + \textit{indel cost} \end{cases}$$

Starting from the lower right cell, every non-decreasing path towards the upper left cell in the comparison table represents an alignment of the two sequences, whereby the best alignment is composed of optimal subalignments. There is always at least one valid path available, yet in most cases there will be several (Needleman & Wunsch, 1970).

Figure 3-12 demonstrates the computation of the minimum edit distance for the strings [BILLIARD] and [BILJART] with one possible alignment, visualised by the yellow path. When applying a constant indel cost of two and a constant substitution cost of one, for the ease of understanding, the minimum edit distance to transform [BILLIARD] into its Dutch equivalent (and vice versa) equals 4. Note that this alignment consists of the same series of actions (insertion, deletion and substitution) as the alignment presented in Figure 3-11. Because the first three letters ("BIL") are identical for both words, the optimal alignment follows a diagonal path down from the top left cell. Next, as the fourth characters ("L" and "J" respectively) are different, the cell diagonally below indicates that a substitution needs to take place, with a cost of 1. In the optimal alignment, however, an indel operation is performed at this moment, shown by the horizontal movement of the yellow path to the right. Indeed, by performing this indel, the future alignment costs are reduced to two substitutions, resulting in a minimum edit distance of 4.

When three or more sequences need to be aligned, a series of pairwise alignments is performed. This method is often referred to as progressive alignment (Wilson, 2006) and results in a series of pairwise distance values between the different sequences, which can be noted in the form of a distance matrix. Figure 3-13 contains a distance matrix representing the minimum edit distances between different translations of the word [Billiard] in respectively English (*EN*), Italian (*IT*), Dutch (*NL*) and Turkish *(TR)*. They were calculated according to the Needleman & Wunsch method, with a constant indel cost of two and a constant substitution cost of one, similar to the example presented in Figure 3-12.

SOURCE

TARGET

		B	I	L	L	I	A	R	D
	0	2	4	6	8	10	12	14	16
B	2	0	2	4	6	8	10	12	14
I	4	2	0	2	4	6	8	10	12
L	6	4	2	0	2	4	6	8	10
J	8	6	4	2	1	3	5	7	9
A	10	8	6	4	3	2	3	5	7
R	12	10	8	6	5	4	3	3	5
T	14	12	10	8	7	6	5	4	4

Fig. 3-12. Computation of the minimum edit distance for the strings [BILLIARD] and [BILJART] and one possible alignment (yellow).

	Billiard *[EN]*	Biliardo *[IT]*	Biljart *[NL]*	Bilardo *[TR]*
Billiard *[EN]*	0	4	4	6
Biliardo *[IT]*	4	0	4	2
Biljart *[NL]*	4	4	0	4
Bilardo *[TR]*	6	2	4	0

Fig. 3-13. The distance matrix containing the minimum edit distance between pairs formed by the strings [Billiard], [Biliardo], [Biljart] and [Bilardo].

Ever since the introduction of the edit distance, multiple variants have been introduced, such as the Edit Distance on Real Sequences (EDR), which calculates the edit distance based on the Euclidean distance (Chen et al., 2005), and, more specifically for the comparison of trajectories, the Normalized Weighted Edit Distance (NWED; Dodge *et al.*, 2012), which takes speed and acceleration of objects into account. In a similar endeavour to find the optimal alignment for sequences that represent movements differing in speed, Dynamic Time Warping (DTW) was introduced by Keogh & Pazzani (2000).

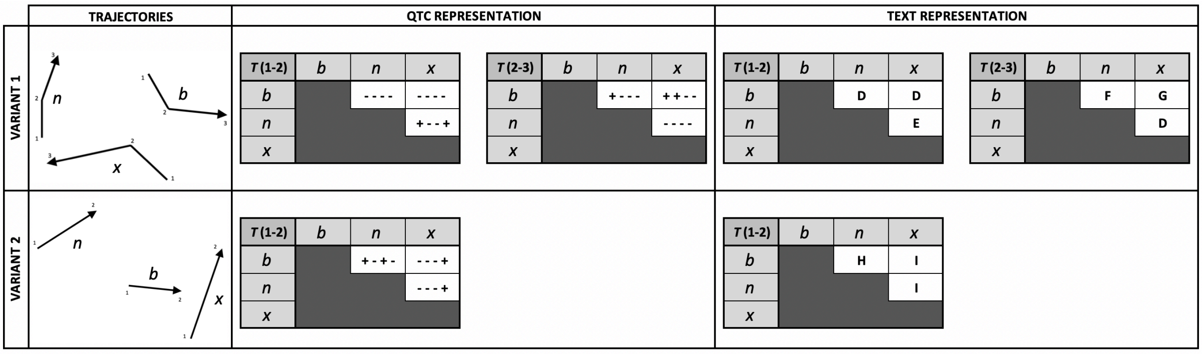
3.7.2 Sequence alignment for matrix sequences

In the previous section, an introduction to basic sequence alignment methods was given. Often, these methods assume letters or other non-divisible elements as the building blocks of the sequences. In biology, DNA sequences are represented by sequences of letters and alignment between them is thus done on sequences of letters. In other fields, such as sociology, letters are often used to represent sociological phenomena. Sequence alignment as such was first implemented in sociology by Abbott (1995), for the analysis of career paths. This triggered a general interest in SAM techniques by sociologists (Abbot & Tsay, 2000), where sequence alignment was seen as a promising methodology for analysing the sequential aspect of human activities in space and time (Delafontaine & Chavoshi *et al.*, 2012). On the crossroads of sociology and geography, for example, Shoval and Isaacson (2007) used the Levenshtein distance metric for sequence alignment of human activity, while Kwan *et al.* (2014) extended this method to include multiple dimensions of human activity. Stehle & Peuquet (2015) extended the basic sequence alignment to align partial patterns using temporal intervals and fuzzy matching, exemplified by the alignment of political transitions of different states during the Arab Spring. For some phenomena, such as the QTC-representation of multiple MPOs as presented in Section 3.1, it is less straightforward to use a character sequence representation. In comparison with existing methods to compare spatio-temporal trajectories of single MPOs based on sequence alignment techniques (Dodge *et al.*, 2012), the comparison of different QTC-matrix sequences can be seen as an alignment of two or more trajectories of multiple MPOs. In this section, we will present methods and techniques that are available to apply SAM on matrix sequences, thus providing distances for a similarity analysis of a number of matrix sequences.

The most straightforward solution for applying SAM to matrix sequences is by replacing the whole matrix by another qualitative representation such as a letter. In that case, a conceptual distance table, containing all distances between all the different (possible) representations would be needed, which might pose a problem when working with huge matrices containing a lot of elements, necessitating huge calculation power. In the rather basic but common case of analysing movement relations between six objects in QTC_{C21}, there are 81^{15} unique QTC-matrices. A bigger problem is that the matrices would be seen as a whole, *i.e.* as elementary building blocks of the sequences, and thus no sequence alignment would be applied on the inner elements in the matrix.

A different solution is to transform the matrix representations to text or other qualitative representations by taking every single element of the matrix as a base. In other words, the sequences would be stripped of their matrix representation, allowing to form sequences containing a series of singular elements which were once the elements that could be found in the cells of the different matrices. When applying this method, basic sequence alignments methods could be applied to the sequences.

Depending on the number of matrices in each sequence, the letter representation would differ in length, yet this would not pose a problem for basic SAMs as argued in Section 3.2.1. An example of this method is given in Figure 3-14, where two movements of MPOs are shown. On the right of these visualisations, QTC_{C21}-matrix representations for both movements are shown, containing a different number of matrices due to the difference in temporal duration. The matrix representations are being transformed by replacing all matrix elements by letters (text representation). The actual alignment according to this method is shown at the bottom of Figure 3-14, where the two letter representations are being aligned, which can be done by the classical SAM methods mentioned earlier.



Alignment						
	D	D	E	F	G	D
	H	I	I			

Fig. 3-14. Letter representation and alignment of two QTC_{C21}-matrix sequences.

There is, however, a downside to this approach, because ignoring the matrix representation may cause some of the valuable information to get lost. The fact that information is stored in matrices means that there is more semantic information included. Corresponding matrices store information about the same timestamps, such as matrices *T*(1-2) of variant 1 and *T*(1-2) of variant 2 in Figure 3-14. Corresponding cells in different matrices store relations between the same objects, such as the cells containing the QTC_{C21}-relation between object *b* and *x* in matrices *T*(1-2) in variant 1 and *T*(1-2) in variant 2. If the different matrix sequences have identical lengths, an alternative for SAM is the pairwise comparison of the elements in the matrix cells.

As such, a distance can be calculated if a conceptual distance table for the possible elements in the matrix is available.

Based on the previous section, it can be concluded that there is currently no optimal sequence alignment method for matrix sequences with different length. Therefore, a novel method for distance calculation between QTC-matrix sequences of different length is suggested in the next section.

3.8 Proposed methodology for distance calculation

In this section, a novel method for distance calculation between matrix sequences, based on matrix sequence alignment, will be presented. First, this will be done for the general case of matrix sequences, regardless of the qualitative calculus used in the matrices. Later, the proposed methodology will be applied more in detail to QTC-matrix sequences of different length. After the application on QTC, a brief reflection regarding other qualitative calculi that could be used with this methodology, is presented.

3.8.1 Levenshtein distance calculation for matrix sequences

As mentioned before, basic sequence alignment methods need to be enhanced in order to compare matrix sequences of different length. New methods should thereby take the special features of matrices into account, such as dependence between cells from different matrices with similar column and row index and temporal relations between matrices located at the same position in different matrix sequences. The proposed method takes the Levenshtein (1966) and the Needleman & Wunch (1970) methods as starting point for the distance calculation. Compared to the basic Levenshtein approach as presented earlier, there are some major differences in the novel methodology.

In the basic sequence alignment methods, two words can be compared and aligned by substituting, deleting and inserting letters. In that case, the words are the sequences and the letters can be seen as the sequence elements. In the novel approach matrix sequences are aligned, where the full matrix sequences can be seen as the sequences and the individual matrices can be seen as sequence elements (Figure 3-15). By this approach, the non-divisible character of the sequence elements, as described in Section 3.2.2, is being violated in the proposed methodology. For the ease of comparison though, we will further use the term 'sequence elements' in the remainder of this paper for both cases.

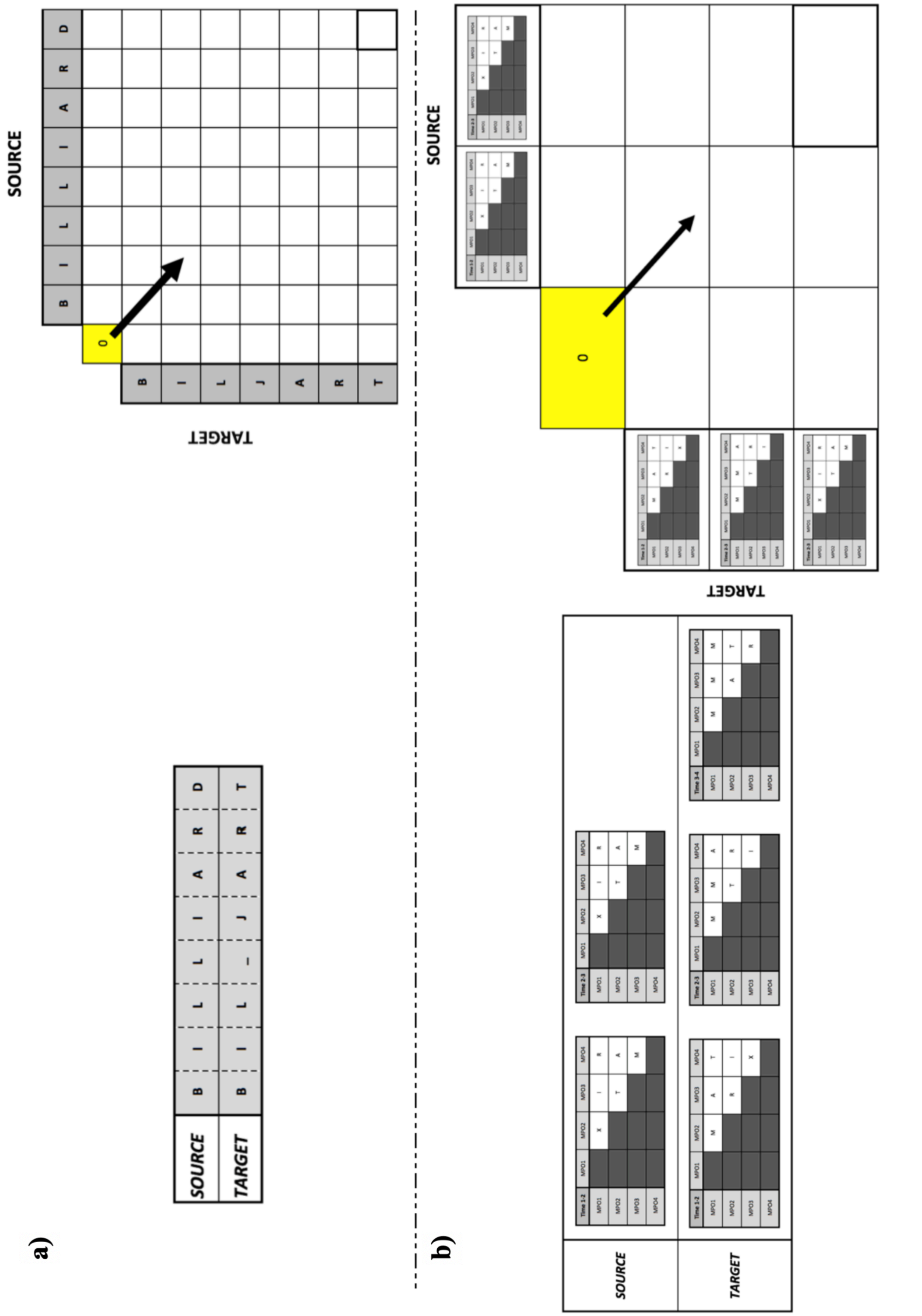


Fig. 3-15. Basic alignment of two strings (a) and novel methodology for alignment of two matrix sequences of different length (b).

If, when comparing two matrices, all cells of the matrices containing the characters of the calculus, are identical, the substitution cost is zero. If some of the cells are not identical, the distance between both matrices can be calculated, in the most basic version of the method based on pairwise comparison of the characters. Dependent on the calculus used to describe movement between the MPOs, this can be done 'on-the-fly' by calculating the distance between every pair of non-identical characters upon occurrence or all the possible transitions between characters can be calculated in advance and saved in a conceptual distance table. It depends on the number of possible transitions (combination of different characters) whether it is more feasible to calculate this on-the-fly rather than looking it up from a pre-constructed conceptual distance table. The indel cost of a whole matrix is equal to the maximum possible substitution cost between two matrices.

Our novel method imposes a restriction on the matrix dimensions, demanding that both matrix sequences should contain matrices having the same dimensions. Because of that, it is possible to perform the rather basic approach of pairwise comparison, since each matrix contains the same number of elements. To fine-tune the distance calculation, matrices can be compared by applying the Levenshtein method on the letter representation of their elements. Based on the preferences of the user and the (type of) sport, by applying different values for the indel and substitution costs, countless variations of distance calculation are possible. One could argue that by applying an infinite indel cost on the internal Levenshtein comparison of the matrices, thereby disabling the deletion and insertion options in the alignment, the method would reduce to the pairwise comparison mentioned earlier. Thus the pairwise comparison of two matrices can be seen as a Levenshtein-comparison with an infinite indel cost.

Where the basic Levenshtein algorithm (see Section 3.7.1) enables to align and provide a distance value for words with a different number of letters, the novel method can provide an alignment for matrix sequences with a different number of matrices, thereby incorporating the special relations (temporal, geographical) that are linked to this matrix representation. Weights for differences between the characters that compose the matrices can be chosen freely for every calculus and can be changed for different applications of every calculus. The weights can be derived from a conceptual distance table.

3.8.2 QTC-matrix sequences

As presented in Section 3.1, QTC-matrices can be used to store the qualitative representations of multiple MPOs through time. A sequence alignment method for this kind of data enables distance calculation between different intervals of moving objects and thus measure the similarity between them. QTC in combination with SAM has already been used to analyse similarity between different human-robot interactions (Hanheide *et al.*, 2012) and movement patterns of body parts of samba dancers (Chavoshi *et al.*, 2015). An important difference, however, is that these approaches cannot deal with interactions between multiple MPOs since the QTC-representations are analysed in specific pairs represented by separate QTC-sequences. As such, the methods proposed by Chavoshi and Hanheide can be seen as a special (more limited) case of the more general method that is presented in the remainder of this chapter. This method is much needed, because often relations between more than two MPOs are described by means of QTC-matrices (Delafontaine & Chavosi *et al.*, 2012). Moreover, QTC-matrices tend to describe movements of MPOs more precisely, since for one MPO multiple QTC-descriptions are given. The combination of these constraints on the movement of that MPO results in a more precise description of its movement.

Two QTC-matrix sequences can be aligned as shown in Figure 3-15b. Each cell of these matrices contains the QTC-relation between the corresponding objects indicated in the row and column header (Figure 3-1). Aligning QTC-matrices requires a conceptual distance matrix for scoring distance between different QTC-characters, displayed in Figure 3-16. In this case, the conceptual distance is a similarity measure for two QTC-relations and is calculated by counting the number of changes in the symbols of the QTC-representations (Van de Weghe & De Maeyer, 2005). When two QTC-characters are identical, the conceptual distance is zero, the conceptual distance between ‘0’ and ‘-’ or ‘+’ is one. The conceptual distance between ‘+’ and ‘-’ is two because direct transition from '+' to '-' or vice versa is impossible due to (theoretical) continuity of the movement (Galton, 2000). The total conceptual distance between two QTC-relations can be calculated by summing up the conceptual distances over all relation characters (Chavoshi *et al.*, 2015). When using qualitative calculi, summing up the conceptual distances will always result in a finite overall distance value. This is important because the indel cost was defined as the maximum possible substitution cost for the whole matrix (see Section 3.3.1). Using a quantitative calculus could result in an infinite indel cost, impeding the use of indels when aligning two matrix sequences, thereby not guaranteeing optimal alignment.

QTC Character	-	0	+
-	0	1	2
0	1	0	1
+	2	1	0

Fig. 3-16. Conceptual distance matrix for QTC-characters.

The Levenshtein distance calculation is applied to the QTC-matrix sequences (Figure 3-17a), including conceptual distance calculations between each pair of individual matrices. These conceptual distance calculations, with one of the calculations shown in Figure 3-17b, provide the distances used for calculating the minimum edit distance for the alignment of the two matrix sequences (Figure 3-17a). Note that by using the conceptual distance calculation presented in Figure 3-17b, a one-on-one identity of the QTC_{C21}-characters of the different matrices is needed for a distance of zero, taking into account the underlying temporal and geographical dependencies of cells of corresponding matrices or cells with identical row and column indices. When multiple static points are included in two QTC-matrix sequences that are compared, the invariant QTC-relations describing the relative movement between the static points, should be omitted from the distance calculation (see Section 3.3).

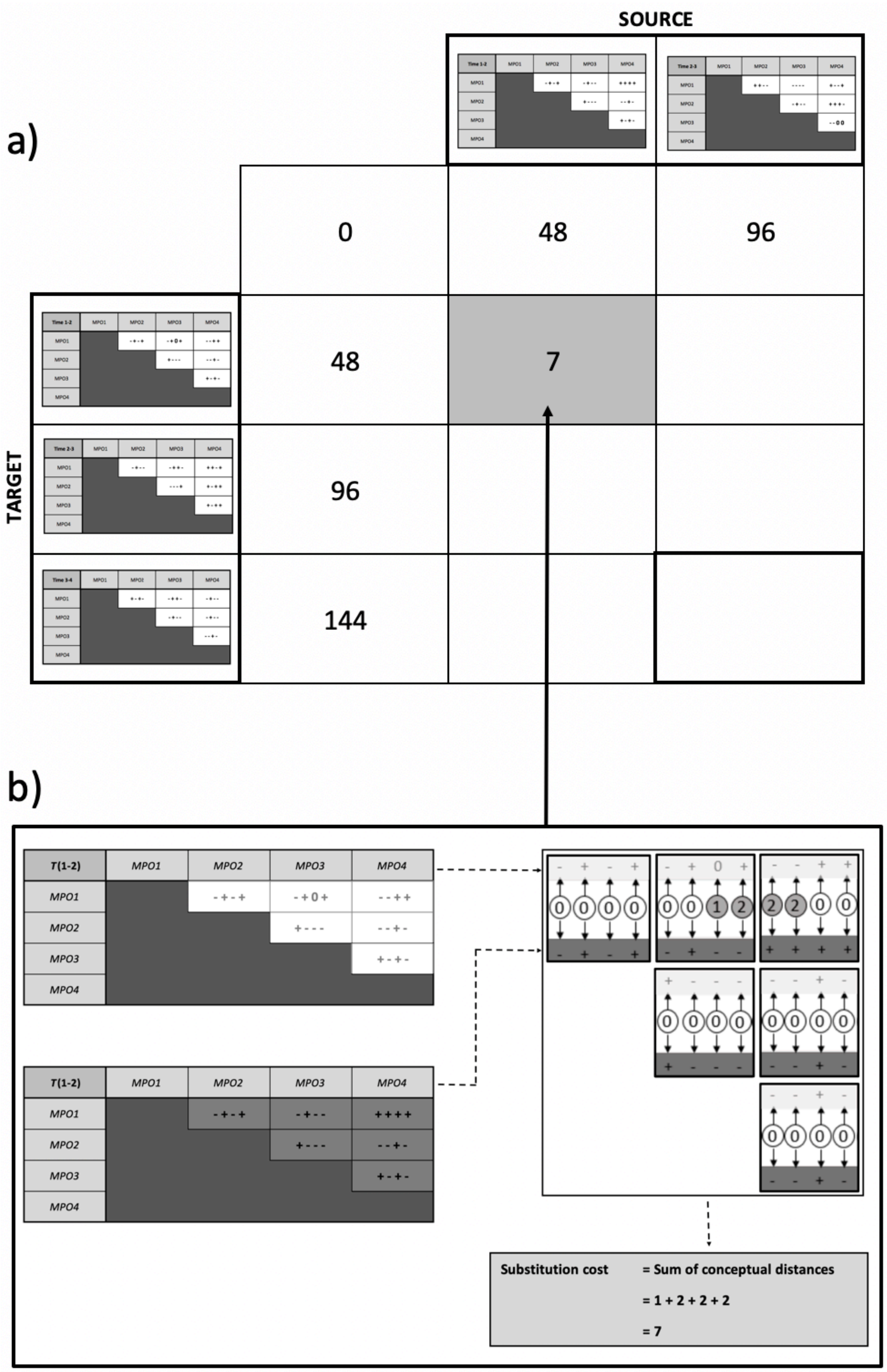


Fig. 3-17. Levenshtein distance calculation for QTC-matrix sequences of different length.

3.8.3 Dynamic three-cushion billiards

In this sports example, the use of the Levenshtein distance calculation for matrix sequences will be demonstrated by performing a spatio-temporal comparison of a series of three-cushion billiards shots. The complete three-cushion billiards dataset containing 55 opening shots, that was presented in Section 3.2, is used for this example.

For performing a spatio-temporal analysis on this dataset, each opening shot can be described with the Qualitative Trajectory Calculus, by means of a sequence of QTC_{C21}-matrices. This QTC-variant was chosen for reasons mentioned in Section 3.4. The QTC_{C21}-representations can then be compared in order to analyse similarities between different shots, which might give an insight in the requirements to perform a successful opening shot or even to facilitate automated recognition of shot types.

Because of differences in how hard the cue ball is played, small variations in where the cue ball hits the edge of the field, or even different playing styles including the addition of ‘effect’ to the cue ball, shots of the same type do not always take the same amount of time. When applying a constant temporal resolution of 20Hz, because of these small differences in temporal length, the QTC_{C21}-representation of the different shots sometimes thus consists of a slightly different number of QTC_{C21}-matrices (fourth column of Figure 3-3). Since traditional methods do not allow calculation of distances between matrix sequences with (slightly) different lengths, this is a perfect sports case to test the novel distance calculation method.

One might propose a transformation from the equal-interval-based dataset to an event-based logging of the coordinates to cope with the differences in temporal length. However, only logging coordinates at specific points, e.g. when one of the balls hits one of the sides of the field, would not result in shots with identical lengths, as different numbers of those events occur in the different shots. It should be emphasized that specific types of shots, such as a successful opening shot, can be achieved in multiple ways. It is for example also possible to first hit the yellow object ball instead of the red one (as was done in this example) to perform a successful opening shot.

A distance matrix containing the distances between each of the shots is calculated by pairwise distance calculation between each combination of two shots. Each pair of QTC_{C21}-matrix sequences is compared using the Levenshtein distance metric, allowing to detect similar movement patterns with (minor) temporal differences. The extent of the temporal differences that can be allowed for detecting movement patterns of different shots, is influenced by multiple factors, e.g. the overall similarity of the movements of the two shots, their lengths (the indel of one matrix will have a smaller impact on the resulting distance when comparing longer shots), the temporal resolution of the data or the distance threshold one defines for labelling two shots as similar. Based on this distance matrix, a hierarchical clustering (Figure 3-18) was performed using the UPGMA (Unweighted Pair Group Method with Arithmetic Mean) method (Sokal & Michener, 1958). In the right part of the figure, it can be seen that, with some explainable outliers, all the shots are clustered according to their type. On the left of the hierarchical clustering, the trajectories of the different shots in those clusters are visualised in an overlay, grouped by the type of shot. Different from the visualisations earlier this chapter, the table background colour was set to light green, to ensure good visualisation of the multiple trajectories which were placed in overlay.

As the validation of these results is done by visual comparison of the ground truth (the type of the shot) with the hierarchical clustering, the trajectories of the two outliers of the clustering are shown in Figure 3-18. In Figure 3-19, the shots ‘6 S’ and ‘52 S’, who are clustered as outliers with the ‘R’ and ‘L’ types respectively, are shown. Although ‘6 S’ is a successful shot (both the red and the yellow ball are hit, and the white ball hits three sides), it can be seen that the trajectories of the balls much more match the other shots of the ‘R’ type. The trajectories of the red and white balls in shot ‘52 S’, on the other hand, much more match the trajectories of the balls of the ‘L’ type than those of the other successful shots. Compared to those shots, the yellow ball shows a rather limited displacement in the shot ‘52 S’. It can be concluded that the hierarchical clustering indeed clusters the shots correctly, based on the trajectories of the balls in those shots. We are convinced that in this way, we supported the use of the proposed method for movement pattern detection of fragments with different lengths, containing movement of multiple MPOs.

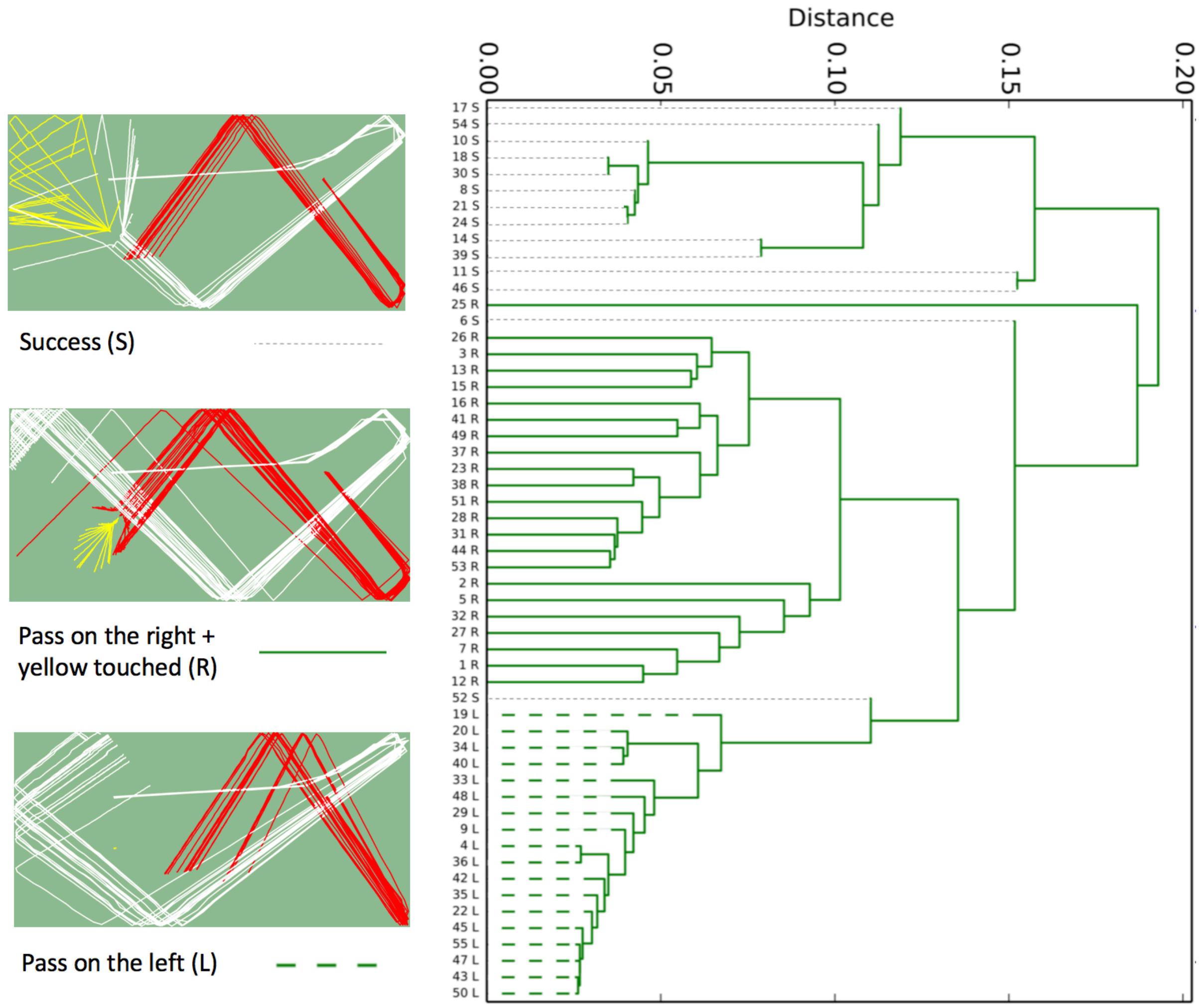


Fig. 3-18. Hierarchical clustering (UPGMA; right) and visualisation (left) of different opening shots by comparing their QTCC21-matrix sequence representations (Success = fine dotted line, Pass on the left = dotted line, Pass on the right & yellow touched = full line).

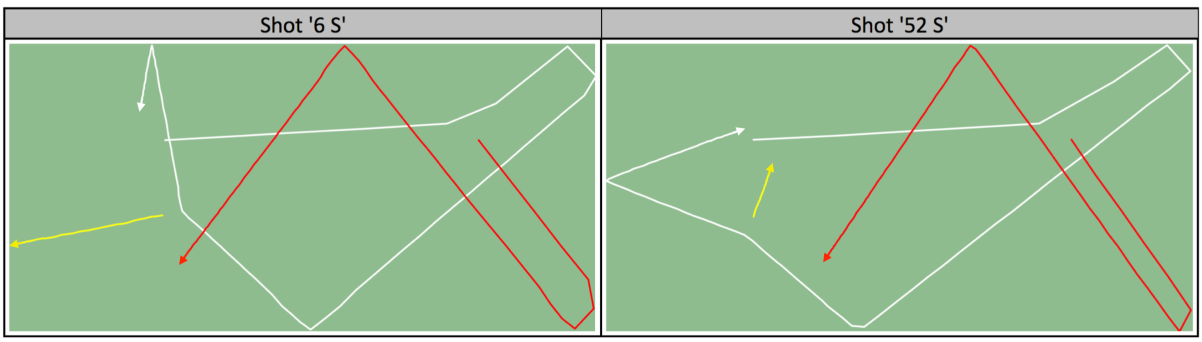
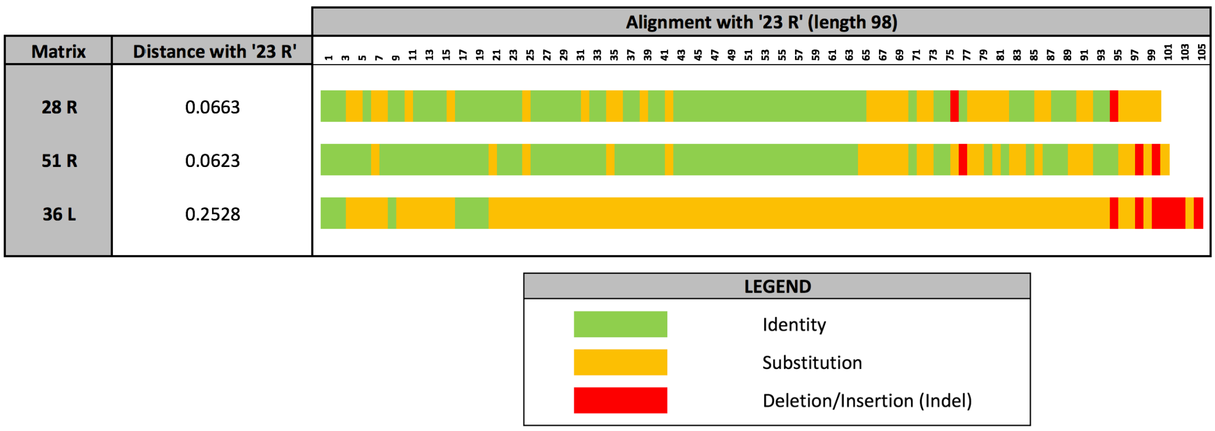


Fig. 3-19. Visualisations of the two shots '6 S' and '52 S' which were clustered as outliers.

To illustrate the alignment of the shots, the alignment and the distances of three shots ('28 R', '51 R' and '36 L') with the shot '23 R' are presented in Figure 3-20. It can be noted that shot '51 R' (represented by 101 QTC_C-matrices) is (slightly) more similar with shot '23 R' (98 QTC_C-matrices) than shot '28 R', which has a more similar number of QTC_C-matrices (100). Despite the higher number of deletions in the alignment of '51 R', it has a lower distance to '23 R' due to a higher number of identical QTC_C-matrices and less costly substitutions. The Levenshtein distance calculation method thus does not only allow to calculate distances between matrix sequences of different lengths, it also allows detection of similar spatio-temporal patterns that happen with different speeds. Since both '28 R' and '51 R' have relatively low distances to '23R', a third alignment ('36 L') with a larger distance is included in Figure 13. Because most of the shots in the dataset start in the same way but tend to differ at the end, most of the insertions and substitutions can be found at the end of the alignments.



3.8.4 Applications of the proposed method

The method presented in Section 3.8 allows for the calculation of distances between matrix sequences of different length, as was shown for the three-cushion billiards. In this dissertation, the method is used for sports analytics of more complex sports in the following chapters. Consequently, the method was implemented in the (prototype) Python program for analysing dynamic sports with QTC. For the distance calculation in the program, both the Levenshtein distance metric as well as a method called 'simple similarity' were implemented. The first method was explained in the previous section, the latter calculates the distance between two QTC-matrix sequences of identical length by pairwise comparison of the QTC-characters in each of the matrices' cells, thereby using the conceptual distances shown in Figure 3-16. For this, all of the distances (between all of the QTC-characters) between the respective QTC-characters of each sequence are summed and divided by the maximum possible distance between these two QTC-matrix sequences. The maximum possible distance is calculated by multiplying the number of QTC-characters in the sequence by the maximum possible conceptual distance between two QTC-characters (which is equal to 2). The Levenshtein distance metric, however, can be applied on both QTC-matrix sequences of identical length as well as different length, whereas the latter method can only be used for sequences with identical length. When calculating distances between such sequences, however, it can be useful to apply the Levenshtein distance metric, as it will take speed differences into account and produces a more accurate distance value for highly similar looking movement patterns that occur at different speeds.

Moreover, the Levenshtein distance metric can be applied on a broad range of phenomena that can be described by qualitative or quantitative calculi in the form of matrix sequences. The approach, for example, is applicable for comparing phenomena described by topological (qualitative) relations. Geographical relations between two MPOs or geographical regions (Egenhofer, 2010) can be described by means of topological relations (Klippel *et al.*, 2008). In this way, topological relations between MPOs or geographical regions can be noted in a matrix form for one specific timestamp. The movement of multiple objects in space can thus be described in a qualitative manner by a sequence of topological relation matrices, where the evolution of the pairwise topological relations for every combination of MPOs or geographical regions is described. The distance calculation method as proposed in this chapter can then be used for comparing the topological descriptions of movements of multiple MPOs or geographical regions during time intervals of different lengths. A conceptual distance table, containing all pairwise distances between the topological relations, is needed to calculate the indel and substitution costs that are used to compute the minimum edit distance.

An important limitation of the proposed methodology, however, is that it can only be used for comparing two matrix sequences with matrices having identical numbers of rows and columns. This means that the matrices contain relations between identical numbers of objects, thus having an identical number of cells. With the proposed method it is therefore possible to compare different sport events containing identical number of actors, e.g. two attacks in basketball or football containing an identical number of players (but with possible different temporal durations). The comparison of phenomena containing different numbers of MPOs, which are usually more difficult to comprehend by human observers, are as such not supported by the proposed method.

REFERENCES

Abbott, A. (1995). Sequence analysis: new methods for old ideas. *Annual Review Sociology,* 21(1), 93-113.

Abbott, A., Tsay, A. (2000). Sequence analysis and optimal matching methods in Sociology. *Sociological Methods & Research*, 29(1), 3-33.

Byrne, R. (1998). Byrne's New Standard Book of Pool and Billiards. San Diego, CA, U.S.A.: Harcourt.

Chavoshi, S.H., De Baets, B., Neutens T., De Tre, G., Van de Weghe, N. (2015). Exploring dance movement data using sequence alignment methods. *PLOS One,* 10(7), 1-25.

Chen, L., Özsu, M.T., Oria, V. (2005). Robust and fast similarity search for moving object trajectories. In Proceedings of the ACM SIGMOD International Conference on Management of Data (pp. 491-502). Baltimore, MD, U.S.A.: ACM.

Cohen, G.L. (2002). Three Cushion Billiards: Notes on the diamond system. *Sports Engineering*, 5(1), 43-51.

Delafontaine, M., Chavoshi, S.H., Cohn, A., Van de Weghe, N. (2012). A Qualitative Trajectory Calculus to reason about moving objects. In M.S. Hazarika (Ed) Qualitative Spatio-Temporal Representation and Reasoning: Trends and Future Directions (pp. 148-168). Hershey, Canada: IGI Global.

Delafontaine, M., Versichele, M., Neutens, T., & Van de Weghe, N. (2012). Analysing spatiotemporal sequences in Bluetooth tracking data. *Applied Geography*, 34, 659-668.

Dodge, S., Laube, P., Weibel, R. (2012). Movement similarity assessment using symbolic representation of trajectories. *International Journal of Geographical Information Science*, 26(9), 1563-1588.

Egenhofer, M.J. (2010). The Family of Conceptual Neighborhood Graphs for Region-Region Relations. *Geographic Information Science,* 6292, 42-55.

Galton, A. (2000). Qualitative spatial change. Oxford, U.K.: Oxford University Press

Gentle, J.E. (2007). Matrix algebra: Theory, computations, and applications in statistics. Berlin, Germany: Springer.

Hanheide, M., Peters, A., Bellotto, N. (2012). Analysis of human-robot spatial behaviour applying a qualitative trajectory calculus. In Proceedings of the 21st IEEE International Symposium on *Robot and Human* Interactive Communication (pp. 689-694). Paris, France: Curran Associates.

Keogh, E., Pazzani, M. (2000). Scaling up dynamic time warping for datamining applications. In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 285-289). New York, NY, U.S.A.: ACM.

Klippel, A., Worboys, M., Duckham, M. (2008). Identifying factors of geographic event conceptualization. *International Journal of Geographical Information Science*, 22(2), 183-204.

Kwan, M.P., Xiao, N., Ding, G. (2014). Assessing activity pattern similarity with multidimensional sequence alignment based on multiobjective optimization evolutionary algorithm. *Geographical Analysis*, 46(3), 297-320.

Levenshtein, V.I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8), 707-710.

Morisson, D.A. (2010). Sequence alignment: methods, models, concepts, and strategies. *Systematic Biology*, 59, 363-365.

Needleman, S.B., Wunch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3), 443-453.

Ranacher, P., Tzavella, K. (2017). How to Compare Movement? A Review of Physical Movement Similarity Measures in Geographic Information Science and beyond. *Cartography and Geographic Information Science,* 41(3), 286-307.

Rosenberg, M.S. (2009). Sequence Alignment: Methods, Models, Concepts, and Strategies. Berkeley, CA, U.S.A.: University of California Press.

Schlich, R. (2003). Homogenous groups of travellers. In: Proceedings of the 10th International Conference on Travel Behaviour Research (pp. 10-15). Lucerne, Switzerland.

Shoval, N., Isaacson, M. (2007). Sequence alignment as a method for human activity analysis in space and time. *Annals of the Association of American Geographers*, 97(2), 282-297.

Sokal, R., Michener, C. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38, 1409-1438.

Stehle, S., Peuquet D.J. (2015). Analyzing spatio-temporal patterns and their evolution via sequence alignment. *Spatial Cognition & Computation: An Interdisciplinary Journal*, 15(2), 68-85.

Tönges, U., Perrey, S.W., Stoye, J., Andreas, W.M.A. (1996). A general method for fast multiple sequence alignment. *Gene*, 172(1), 33-41.

Van de Weghe, N., Cohn, A.G., De Maeyer, P., Witlox, F. (2005). Representing moving objects in computer-based expert systems: the overtake event example. *Expert Systems with Applications*, 29 (4), 977-983.

Van de Weghe, N., De Maeyer, P. (2005). Conceptual neighbourhood diagrams for representing moving objects. *Lecture Notes in Computer Science,* 3770, 228-238.

Waterman, M.S. (1984). General methods of sequence comparison. *Bulletin of Mathematical Biology*, 46(4), 473-500.

Wilson, C. (2006). Reliability of sequence-alignment analysis of social processes: Monte Carlo tests of ClustalG software. *Environment and Planning A*, 38, 187-204.

Wilson, C. (2008). Activity patterns in space and time: calculating representative Hagerstrand trajectories. *Transportation*, 35, 485-499.

4.

QTC FOR STATIC SPORTS ANALYTICS

In this chapter, the proposed methodology for static sports analytics based on QTC_S is introduced. The aim is to use the method for team tactics analysis in team sports. First, the changes to the original QTC_S-methodology are discussed, as argued in the first paragraph of Section 2.3.2 Next, a new and more suitable method for the calculation of distances between QTC_S-representations of different sports point formations is presented. Subsequently, the challenges that are associated with static sports analytics based on QTC_S, are introduced. Since these challenges are merely implementation issues and very similar to the challenges presented in Chapter 3, they are presented rather briefly in this chapter by means of basic three-cushion billiards examples. The three-cushion billiards examples are based on the dataset that was introduced in Section 3.2. The first challenge is the differentiation between sports point formations that occur at different absolute locations, by including static points into the respective QTC_S-descriptions. The second challenge lies in the decision on the thresholds for the assignment of the different QTC_S-characters. The last challenge is the decision on whether to take permutations of the sport point formations into account when calculating distances. A more concrete sport case of the proposed method for static sports analytics is presented in Chapter 8.

4.1 Different approach QTC_S

The sports analytics method presented in this chapter requires a (slightly) different interpretation of the original QTC_S-method that was introduced in 2004 (Van de Weghe, 2004; Van de Weghe *et al.*, 2005). In the original approach, QTC_S is used to describe shapes (polylines and polygons) by means of QTC-characters, rather than dynamic movements of MPOs (see Section 2.2.4). Polylines and polygons consist of points that are connected by vectors in an ordered manner, *i.e.* each of the points making up the polygon or polyline are connected with one or maximum two other fixed points, by means of a vector. The goal of the proposed method, however, is not the description of such shapes, but rather the description and comparison of sports point formations. Examples of such sports point formations are, amongst others, the positions of billiard balls on a billiard table at a specific moment, the positions of football players on the football field at a specific moment or the positions of volleyball players at a specific moment such as the moment of the serve. In the case of describing polylines and polygons consisting of a large number of points with QTC_S (or even infinite numbers in the case of shapes with curved edges), the original method is often preceded by a selection (thus reduction) of the points defining the shape (Van de Weghe *et al.*, 2005). For shapes consisting of a large number of points, this reduction of points reduces the size of the QTC_S-description, while in the case of shapes described by an infinite number of points (e.g. shapes that include curved edges), this selection is a necessity to be able to describe the shape with QTC_S. In static sports analytics, however, these points are given in advance, as the positions of players can be derived from sports data. Although this difference is important, it mainly reduces the required processing time for calculating the QTC_S-description in sports, by eliminating the first step. The difference in methodology, however, lies in the fact that these sports data are point formations rather than a polyline or polygon. Because of this, there is no defined ordering of the points and therefore no definition of the vectors between the different points making up the point formation. The difference in the interpretation of QTC_S therefore is that the relations between all of the possible vectors between all point combinations need to be described for static sports analytics. This difference is illustrated by a simple example in Figure 4-1. On the left, the original approach is shown by the construction of four vectors between the points defining the shape. The construction of the vectors of the different approach is shown on the right, for the exact same points. Consequently, given a fixed number of points for both approaches, the QTC_S-matrix of the sports analytics method will have bigger dimensions than for the original approach (see Figure 4-2). Note that the polyline shown in this figure is only one of the possible polylines that could be constructed between the three billiard balls.

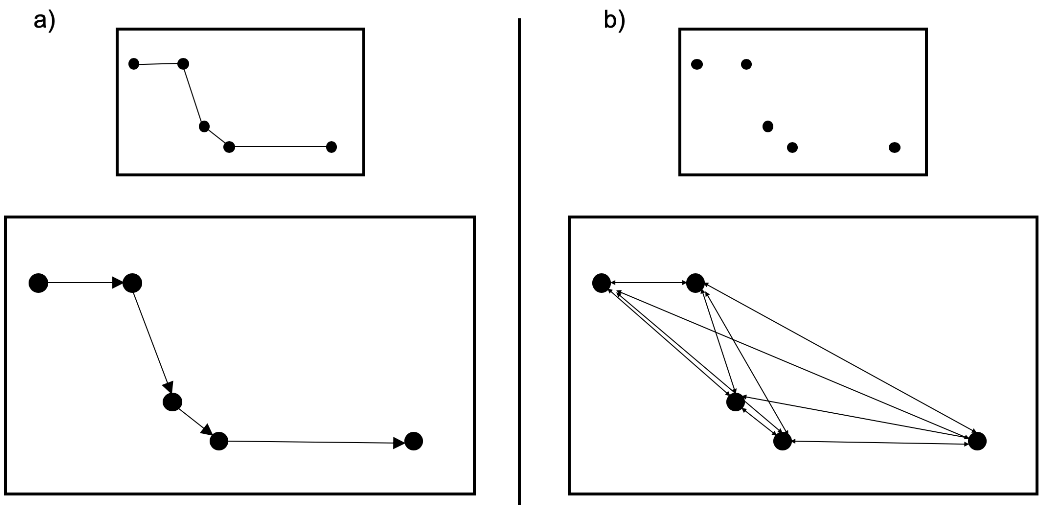
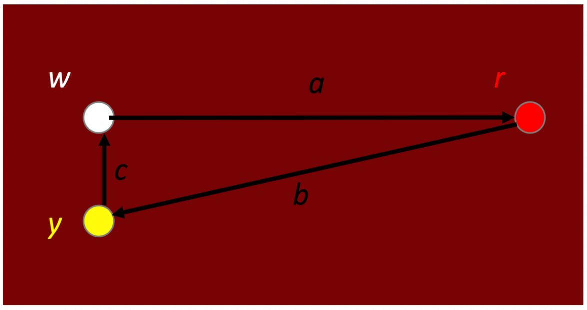


Fig. 4-1. Difference in approach of the original QTC_S method (a) and the QTC_S method for static sports analytics (b). Note that in the latter, the vectors have double directions.

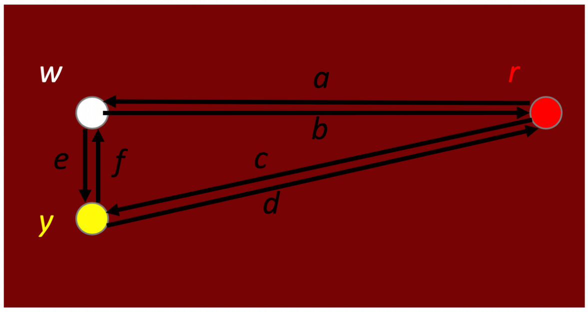
The fact that all possible vectors between all of the point combinations are described has some important consequences. An advantage is the more detailed description of a point formation, although the risk of over-averaging significant differences between two point formations might be higher. Also, more computation power will be required to calculate and compare QTC_S-descriptions of point patterns. An advantage, however, is that the more detailed description makes it feasible to describe the relation between two vectors by means of a level one QTC_S-relation consisting of only two characters (derived from level one QTC_B), thereby reducing the size of a QTC_S-description. To include more detail into the description, level two QTC_S can be used, which is derived from the level two QTC_B. For the original QTC_S-method, the QTC_S-matrix contained relatively more cells where vectors with identical start or endpoints are described. For such cases it is true that QTC_B might give a description with too little detail as some of the characters are not possible to occur. While in the original approach four characters are used (derived from level one QTC_C), the two or three-character variants (level one and level two QTC_S) seem better suited for static sports analyses, because these reduce the need for computing power, which is already increased by the higher matrix dimensions. Moreover, using less characters in the QTC_S description reduces the risks and effects of over-averaging in this case. In this dissertation, only the level one variant is used, which is written as QTC_S for convenience.

Since QTC_S for sports analytics does not describe a shape but rather a point formation at a specific timestamp, the method is named 'QTC Static' instead of 'QTC Shape'. The points that are described by QTC_S are by definition point objects (POs) that are not moving, as only one timestamp is taken into account. In sports analytics, however, the goal of the method is to compare the formation of players at different times (by comparing the different QTC_S-matrices at the respective timestamps). The POs that are described with QTC_S are therefore players that do actually move through time. For this reason, it was decided that, in this dissertation, the objects that are described by QTC_S for static sports analytics are named MPOs (Moving Point Objects). The main argument is that the studied objects are the same as for dynamic sports analyses, which will be often conducted alongside the static analysis method. Moreover, one could argue that the static sports analytics could in fact be a dynamic analysis, if every timestamp of a fragment of a dynamic analysis is analysed and compared by means of QTC_S. Although in that case the same timestamps are taken into account as for a dynamical analysis, only the positions of the players at each of the timestamps are taken into account, and not the displacement of the players that happen in between those timestamps. A last but important difference is the fact that, with QTC_S, moments and not fragments are compared, in comparison with dynamic QTC-variants. A 'moment' therefore refers to the spatial configuration of two or more MPOs at one time instance (timestamp).

Original method	Static sports anaytics method



t1	a	b	c
a		- -	0 -
b			- -
c			



t1	a	b	c	d	e	f
a		- -	+ +	- -	- +	- -
b			- -	0 -	+ +	0 -
c				- -	- -	- -
d					- -	+ +
e						- -
f						

Fig. 4-2. Difference in approach of the original QTC$_S$ method (QTC Shape, on the left) and the QTC$_S$ method for static sports analytics (level one QTC Static, on the right), shown with a three-cushion billiards example. Note that the vectors (shown in black) between identical points with a different direction are shown alongside each other only for visualisation purposes.

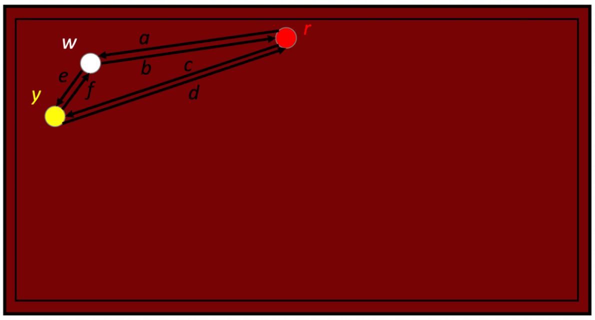
4.2 Distance calculation between different sports point formations

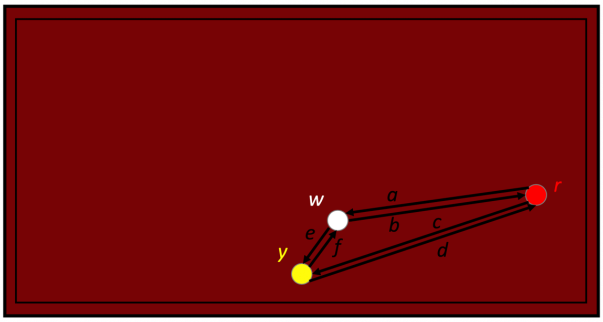
The distance calculation between two QTC$_S$-matrices as described by the original method (Van de Weghe, 2004; Van de Weghe *et al.*, 2005) only counts the relative number of different cells of both matrices, and thus does not take the exact differences of the respective QTC-characters into account. In Van de Weghe & De Maeyer (2005), however, conceptual distances between different QTC-characters are described (see Figure 3-16), which could be used to calculate the distance between two QTC$_S$-matrices more precisely. When two QTC-characters are identical, the conceptual distance is zero, the conceptual distance between '0' and '-' or '+' is one. The conceptual distance between '+' and '-' is two because direct transition from '+' to '-' is impossible in a continuous system (Galton, 2000). As argued in Section 2.3.1, however, sports positional data is not continuous and the intermediate '0'-character is therefore not guaranteed. For static sports analytics this implies that a direct transition from a '-' to a '+' character or vice-versa is possible for an identical QTC-relation between two QTC_S-matrices of consecutive timestamps. Similar to the approach of Chavoshi (Chavoshi *et al.*, 2015), the total conceptual distance between two (level one QTC_S) or three (level two QTC_S) QTC-relations can be calculated by summing up the conceptual distances over all relation characters of the both QTC_S-matrices. Next, this sum of conceptual distances is divided by the maximum possible conceptual distances between both matrices (which can be calculated by using the dimensions of these matrices). It is important to mention that with this method for distance calculation, it is not possible to compare QTC_S-matrices of different dimensions. More concretely, the static sports analytics method is limited to comparing sports point formations of an identical numbers of points, *i.e.* players.

4.3 Using static points

Similar to its dynamic counterpart, QTC_S cannot discern between identical point formations occurring at different absolute locations. Figure 4-3 illustrates the fact that two such point formations result in identical level one QTC_S-descriptions and thus a distance of zero when using the method for distance calculation presented in Section 4.2. The reason for this is that only the relative positions between the different points are taken into account. For a three-cushion billiards player, however, there is an important difference between the configuration of the different billiard balls at $t1$ and $t2$. Assuming moments 1 and 2 display the end positions of a three-cushion billiards opening shot, they impose very different requirements on a billiards player for performing the next shot, because of their different absolute locations (with respect to the table). The reason for this is that the borders of the table are used by the player to deflect the cue ball in the correct manner to perform a successful shot. Similarly, differences in absolute location of sports point formations in other sports might change the context of the formations and could therefore be of importance to a sports analyst. For this reason, it is required to revise the method to allow to detect these differences.

MOMENT 1	MOMENT 2





t1	a	b	c	d	e	f
a		--	++	--	-+	--
b			--	+-	++	+-
c				--	-+	--
d					--	++
e						--
f						

t2	a	b	c	d	e	f
a		--	++	--	-+	--
b			--	+-	++	+-
c				--	-+	--
d					--	++
e						--
f						

Fig. 4-3. Two three-cushion billiard moments containing the positions of three billiard balls *w*, *y* and *r* at two timestamps *t1* and *t2* and their identical level one QTC_S-representations.

By including static points into the QTC_S-descriptions of different sports points formations, a difference in absolute location between two identical formations can result in a different QTC_S-relation of their respective QTC_S-matrices, provided that the static point is chosen in a correct manner to discern between both. A static point in this sense is an identical point that is added to the different sports point formations that are compared using QTC_S. To demonstrate this, Figure 4-4 displays the impact of the addition of a static point *s* on the level one QTC_S-matrices of the sports point formations of Figure 4-3. To ensure the readability of this figure, the vectors between each of the MPOs and static points are shown in a generalized manner, *i.e.* by showing the identical vectors with opposite direction as one. Identical cells of both QTC_S-matrices are shown in green and cells containing one or more different QTC-relations are shown in red. While the addition of a static point in Figure 4-4 results in the inclusion of different QTC_S-relations in the respective QTC_S-matrices, the exact effect is decided by the exact placement of the static point for each application.

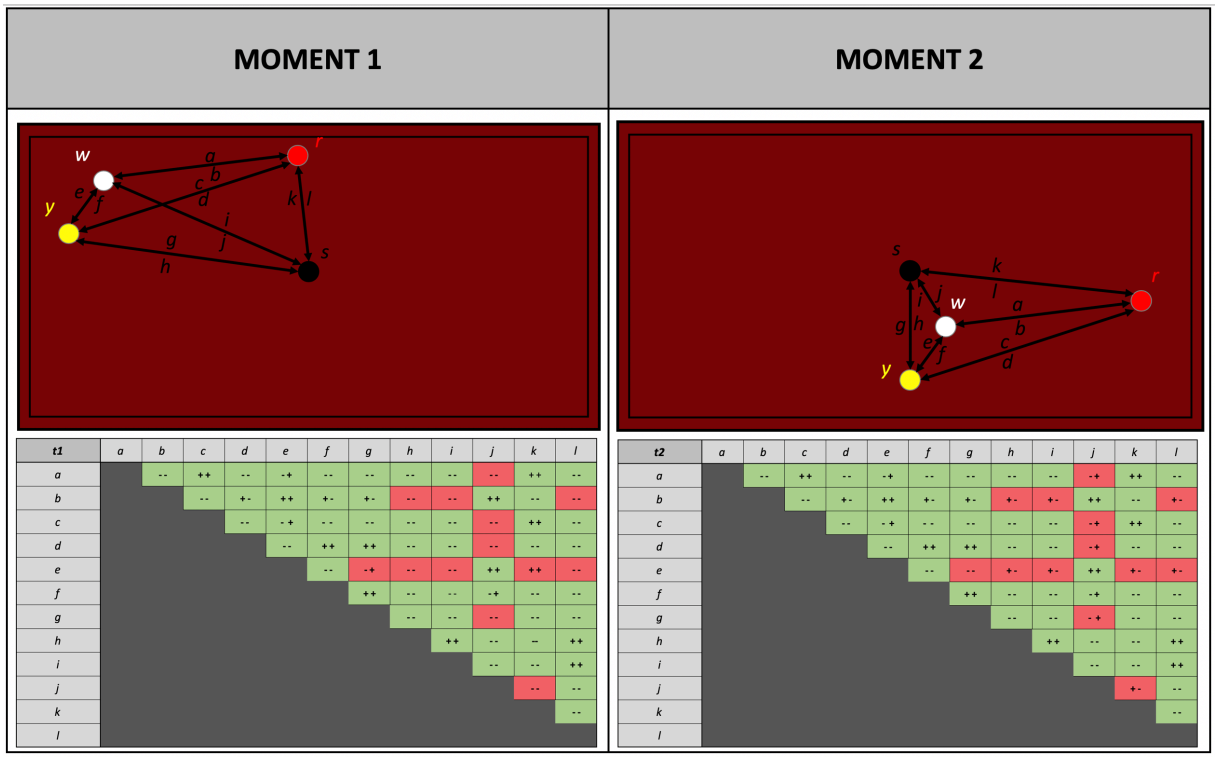


Fig. 4-4. Two three-cushion billiard moments containing the positions of three billiard balls *w*, *y* and *r* and one static point *s* at two timestamps *t1* and *t2* and their non-identical level one QTCS-representations.

In Figure 4-4 it can be observed that the addition of just one static point more than doubles the number of cells of the respective QTCS-matrices. In fact, for every additional (static) point added to a (sports) point formation, the amount of information in the QTCS-matrix increases exponentially, since the number of cells (β) in a QTCS-matrix describing an *n*-point sports point formation can be calculated using following formula[8]:

$$\beta = \frac{(n*(n-1))^2 - n*(n-1)}{2}$$

[8] This formula is valid for all QTCS-descriptions of (sports) point formations, regardless of whether the points are static points or MPOs.

Consequently, the number of cells increases quickly when describing sports point formations of a large number of points, resulting in a high risk of over-averaging small but significant differences between two different sports point formations. There are, however, some practices that can reduce the number of cells that are taken into account while comparing two QTC_S-matrices. When describing a sports point formation including multiple static points, for example, the cells describing QTC-relations between vectors whose starting and endpoints are both static points, should be omitted from the distance calculation (see Section 4.2). Because of the definition of a static point, these relations are identical for all the moments that are compared. In comparison with the dynamic QTC-variants, however, it is not needed to omit half of the QTC-relation between a vector which has a static point for both the start and endpoint and another vector (which is not defined by two static points).

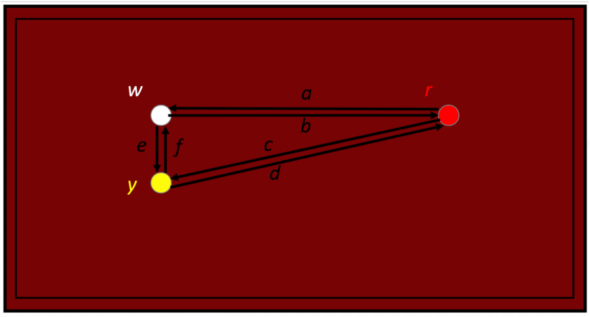
4.4 Using angular tolerances for the assignment of the QTC-characters

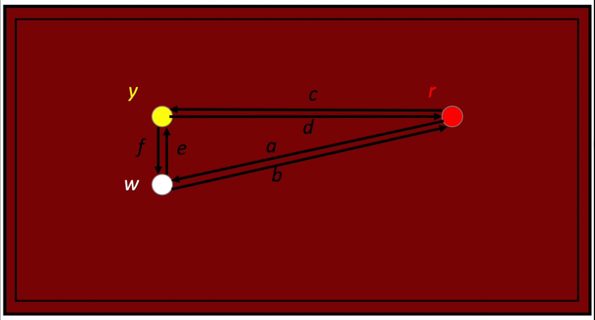
Similar to its dynamic counterpart, QTC_S is characterized by the relatively low chance of occurrence of the 'o'-character in real-world situations, as argued in Section 3.5. Despite the difference of the nature of the described phenomena, *i.e.* formations of players at singular timestamps instead of dynamic movement patterns of players, the manner in which the vectors are described is identical (see Section 4.1). For this reason, the concept of including angular tolerances for the assignment of the 'o'-character might seem to enable the method to produce better results for similarity analyses. Including such an angular tolerance for the 'o'-character, however, often has some unwanted and unexpected ramifications on (other) QTC-relations, as was shown in Section 3.5. Hence, a parameter to define an angular tolerance for the 'o'-character for the calculation of the QTC_S-matrices was included in the Python program (see Chapter 5), but was very rarely used for the case studies conducted in this dissertation.

4.5 Using permutations

Another challenge is to define the importance of roles when comparing different sports point formations. In layman terms this can be described by following question: "Is it important who (e.g. which players, ball, etc.) constitutes to which part of the sports point formation?". More precisely, should two moments be considered as identical when all of the points making up the sports point formations of both moments are identical but if two or more MPOs take up each other's place? An example showing two such moments containing identical sports point formations is shown in Figure 4-5. For both moments the formation and the level one QTC_S-matrices of three billiard balls, *w*, *y* and *r* are shown. The second moment is identical to the first, however with balls *w* and *y* taking each other's positions. Note that the labels of the vectors in Figure 4-5 are defined equally by means of their start and endpoints for both moments. Using the (standard) pairwise comparison of the QTC-relations of both QTC_S-matrices, a distance > 0 will be detected, indicating that both moments are not identical. Taking all possible permutations of both QTC_S-matrices into account when calculating the distance between them, however, the pairwise comparison will result in a distance of zero, implying both moments are identical.

MOMENT 1	MOMENT 2





t1	a	b	c	d	e	f
a		--	++	--	-+	--
b			--	0-	++	0-
c				--	--	--
d					--	++
e						--
f						

t2	a	b	c	d	e	f
a		--	++	--	--	--
b			--	-0	++	--
c				--	--	-0
d					0-	++
e						--
f						

Fig. 4-5. Two billiard moments showing the positions of three billiard balls *w*, *y* and *r* and their level one QTC$_S$-representations, illustrating the problem of permutations for static sports analytics.

The Python program includes a parameter to let the user define whether or not to take these permutations (or a limited list of permutations) into account. Given this, the calculation and comparison of the permutations is a process that requires a lot of computing power. While the calculation of permutations for dynamic sports analytics seems feasible, the dimensions of the QTC$_S$-matrices are larger, even when describing sports point formations with limited numbers of points. From benchmark tests of the Python program, it turned out that it is possible (*i.e.* with acceptable execution times) to use permutations for static sports analytics, if the total number of points (including optional static points) in the sports point formation is not exceeding 4. Table 4-1 presents such a benchmark test, showing run times (s) for the comparison of sports point formations of 5 and 20 different moments with level one QTCs for different numbers of points. This comparison includes the generation of the QTC$_S$-matrices as well as the distance calculation between all of the respective QTC$_S$-matrices (thus calculation of the distance matrix between all of the compared moments). It can be seen that for comparing small numbers of moments, permutations may be used with up to 5 points. In real-world cases where larger numbers of moments are compared, however, the usage of permutations is limited to cases with maximum 4 points.

Table 4-1. Run times (s) of the Python program for various QTC_S analyses, illustrating the exponential effect of permutations on run times.

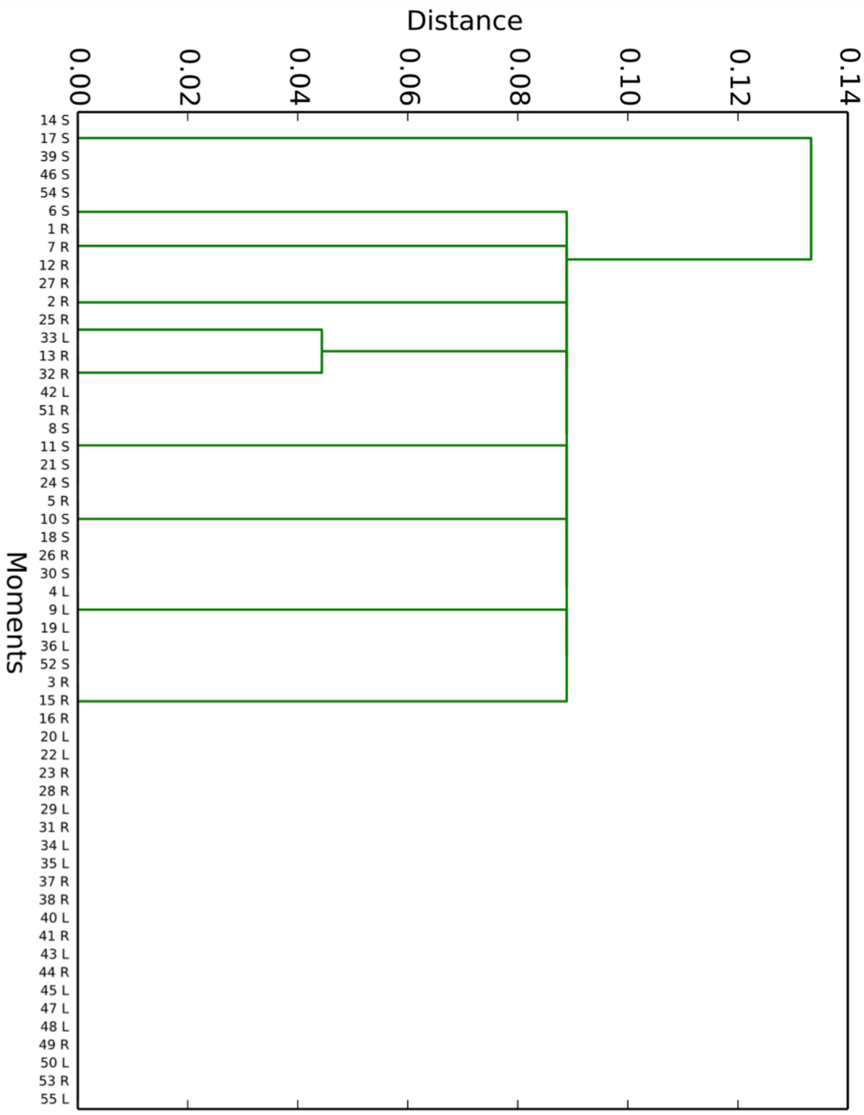
		Number of points				
	Permutations	3	4	5	6	7
5 moments	No	0.95	1.27	1.37	1.50	1.72
	Yes	1.23	1.51	7.35	79.84	1011.83
20 moments	No	3.05	3.68	4.08	4.61	7.22
	Yes	3.15	6.60	53.46	626.08	> 4800

4.6 Static three-cushion billiards example

In this sports example, the static sports analytics method is demonstrated by an analysis of the three-cushion billiards dataset that was presented in Section 3.2. The end positions of the three billiard balls of each of the 55 opening shots are described by means of QTC_S-matrices and the distance between them is calculated. The visualisations of the positions of the balls at the end of each opening shot are included in Appendix 1. As the exact positions of the balls are important for a billiards player to determine the strategy of the next shot, no permutations were used in this example. Because the limited number of points that make up the sports point formations, *i.e.* 3 billiard balls, the most detailed QTC-variant (level two QTC_S) is used. Moreover, since each of the opening shots is performed in a similar manner (see Section 3.8.3), it is to be expected that each of the end situations of the shots will have a high resemblance with each other. For these two reasons it seems more feasible to use the more detailed level two QTC_S in this example. First, the analysis was performed without using any static point(s). Distances between all of the QTC_S-matrices were calculated and summarized by means of a distance matrix. Based on this distance matrix, a hierarchical clustering (Figure 4-6) was performed using the UPGMA (Unweighted Pair Group Method with Arithmetic Mean) method (Sokal & Michener, 1958). To evaluate the hierarchical clustering and to define the important clusters, an inconsistency analysis is performed on the hierarchical clustering tree. In this analysis, the relative consistency of each of the links is expressed in terms of an inconsistency coefficient. This coefficient is calculated by comparing the height of each link with the average heights of the links below it, and is normalized by the standard deviation for the moments included in the cluster at each level. As such, the inconsistency coefficient expresses the discriminating power of each link, where links with higher inconsistency coefficients are more important for differentiating between important clusters than links with lower inconsistency coefficients.

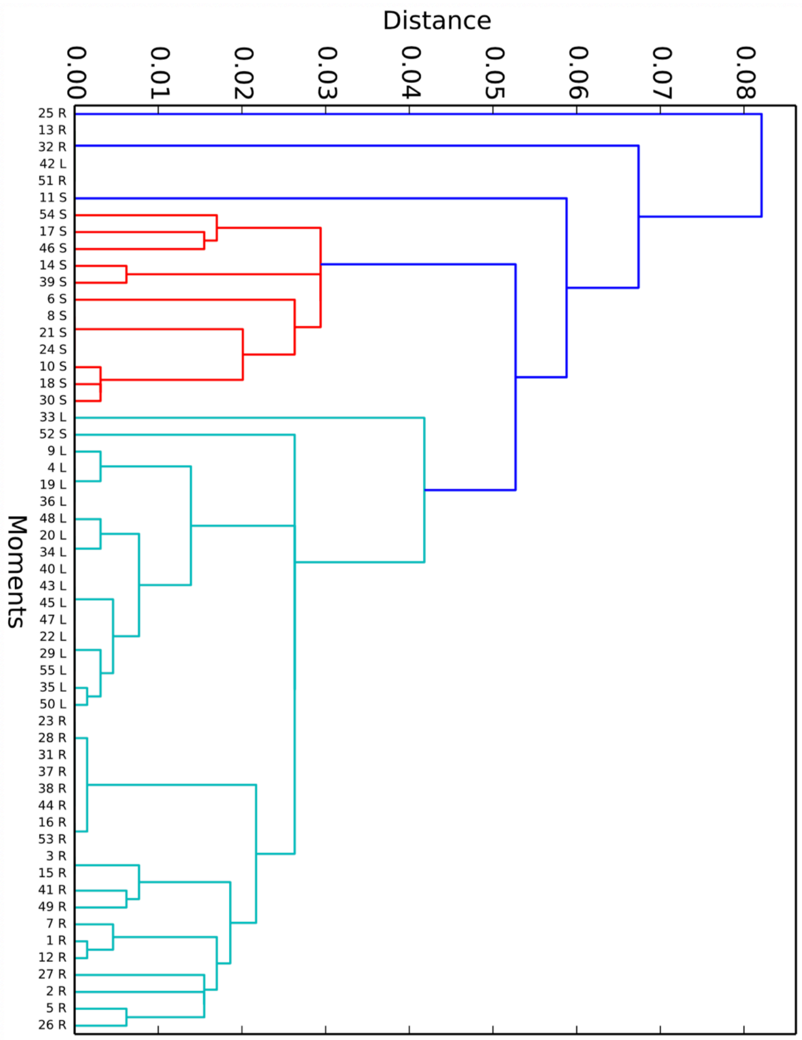
The hierarchical clustering of the example, presented in Figure 4-6, does not show a clear distinction between the different groups of shots (L, R and S) at either of the cluster levels. The possible distances between level two QTC_S-matrices are limited because only three billiard balls are taken into account (which limits the dimensions of the QTC_S-matrices). Moreover, because most of the moments have a distance of zero between them, the hierarchical clustering does not provide much insight in the possible clustering of the different moments. While the inconsistency analysis might support the evidence for 8 or even 9 clusters, the moments included in these clusters do not have resembling properties (*i.e.* do not belong to similar groups of shots).

Based on the remarks in the previous paragraph, one might propose to add four static points, one at each of the corners of the billiards table, to increase the level of detail of the QTC_s-descriptions. Adding these four static points might enable the detection of smaller differences of the positions of the three balls at the different moments, and takes rotation of the point formation into account. As shown in Figure 4-7, this results in more differentiating distance values between the moments and accordingly in a more detailed clustering. However, it can be seen that because of over-averaging, overall distances between the different moments are relatively low. The hierarchical clustering in Figure 4-7 is colour coded into clusters based on the inconsistency analysis. Looking at the tree in a top-down manner, link number four (with an inconsistency coefficient of 0.98) seems to be a first meaningful separation of the moments into a (red) cluster with (almost all of the) successful shots (group 'S') and a (light blue) cluster with mainly unsuccessful shots (groups 'L' and 'R') and some clusters containing (explainable) outliers. Moving down the tree, the cluster with unsuccessful shots is later on divided almost perfectly into two separate clusters containing respectively the moments from the 'L' and the 'R'-shots. Without going into detail on the conclusions and implications of this tree on three-cushion billiard tactics, it clearly demonstrates the abilities of the proposed method for static sports analytics based on QTC_S.



Link	Inconsistency	Link	Inconsistency
1	0.00	28	0.00
2	0.00	29	0.00
3	0.00	30	0.00
4	0.00	31	0.00
5	0.00	32	0.00
6	0.00	33	0.00
7	0.00	34	0.00
8	0.00	35	0.00
9	0.00	36	0.00
10	0.00	37	0.00
11	0.00	38	0.00
12	0.00	39	0.00
13	0.00	40	0.00
14	0.00	41	0.00
15	0.00	42	0.00
16	0.00	43	0.00
17	0.00	44	0.00
18	0.00	45	0.00
19	0.00	46	1.15
20	0.00	47	1.15
21	0.00	48	0.58
22	0.00	49	0.58
23	0.00	50	0.58
24	0.00	51	0.00
25	0.00	52	0.58
26	0.00	53	0.00
27	0.00	54	0.87

Fig. 4-6. Hierarchical clustering (UPGMA) of the end positions of the billiard balls of the 55 different opening shots by comparing their level two QTC_S-matrices (left) and the inconsistency analysis displaying the inconsistency coefficients for each of the links (right).



Link	Inconsistency	Link	Inconsistency
1	0.00	28	0.71
2	0.00	29	0.87
3	0.00	30	0.71
4	0.00	31	0.00
5	0.00	32	0.00
6	0.00	33	0.00
7	0.00	34	1.09
8	0.00	35	0.75
9	0.00	36	1.05
10	0.00	37	0.00
11	0.00	38	0.71
12	0.00	39	0.00
13	0.00	40	0.68
14	0.00	41	0.71
15	0.00	42	0.71
16	0.00	43	1.14
17	0.00	44	0.71
18	0.00	45	0.90
19	0.00	46	0.00
20	0.00	47	0.71
21	1.15	48	0.70
22	0.00	49	0.58
23	0.00	50	0.71
24	1.01	51	0.98
25	0.71	52	0.71
26	0.00	53	0.69
27	0.00	54	0.71

Fig. 4-7. Hierarchical clustering (UPGMA) of the end positions of the billiard balls and 4 static points of the 55 different opening shots by comparing their level two QTC_S-matrices (left) and the inconsistency analysis displaying the inconsistency coefficients for each of the links (right).

REFERENCES

Chavoshi, S.H., De Baets, B., Neutens T., De Tre, G., Van de Weghe, N. (2015). Exploring dance movement data using sequence alignment methods. *PLOS One,* 10(7), 1-25.

Galton, A. (2000). Qualitative spatial change. Oxford, U.K.: Oxford University Press

Sokal, R., Michener, C. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38, 1409-1438.

Van de Weghe, N. (2004). *Representing and Reasoning about Moving Objects: A Qualitative Approach*. Ph.D. dissertation, Ghent University.

Van de Weghe, N., De Maeyer, P. (2005). Conceptual neighbourhood diagrams for representing moving objects. *Lecture Notes in Computer Science,* 3770, 228-238.

Van de Weghe, N., De Tré, G., Kuijpers, B., De Maeyer, P. (2005). The double-cross and the generalization concept as a basis for representing and comparing shapes of polylines. In R. Meersman, Z. Tari, P. Herrero (Eds.), On the move to meaningful internet systems. Lecture Notes in Computer Science (pp.1087). Berlin, Heidelberg, Germany: Springer.

5.

IMPLEMENTATION

In this chapter, the implementation of the dynamic and the static sports analytics methods, which were introduced in Chapters 3 and 4, is discussed. Earlier in this dissertation, this implementation was referred to as 'the Python program'. In fact, the implementation consists of two individual Python programs[9], written specifically for the research presented in this dissertation. In the first section of this chapter, the code of the programs is discussed by means of generalized flowcharts. Next, different implementations of QTC are discussed, together with their implications for real-world (sports) cases.

5.1 Code flowcharts

In this section, the implementation of the dynamic and the static sports analytics methods into two Python programs is discussed. Since a detailed discussion of the code falls outside of the scope of this dissertation, the two programs are presented by means of generalized flowcharts. Flowcharts are excellent for presenting the structure of a process or algorithm in a clear, simplified manner, and, in this case, allow to focus on the essential parts of the Python programs. Flowcharts for both programs (dynamic and static) are presented and discussed separately in the next sections.

[9] Note that, for this research, various other Python programs were written. In this dissertation, only the most important and general programs are discussed, to give the reader an overview of the different steps needed to conduct the presented methods for dynamic and static sports analytics based on QTC.

5.1.1 Dynamic sports analytics

The Python program for the dynamic sports analytics method can be summarized in the generalized flowchart shown in Figure 5-1. In this chart, functions are shown by blue boxes, including a description of the function, and an overview of the input and output. Numbers (show in light blue) indicate the order in which the different functions are called upon by the program. Performing a dynamic sports analysis starts with the user defining the variables mentioned in the 'Main()'-function. Some of these parameters are binary, while others require a more specific input of the user. Default values, if applicable, are mentioned in between brackets. As such, if a user does not specify otherwise, the analysis will be conducted in two dimensions (x,y), with the original temporal resolution[10] of the dataset and thus without angular tolerance or permutations. Furthermore, the Levenshtein distance metric is not selected as default, implying that the analysis is limited to the comparison of fragments with identical temporal lengths. Consequently, the 'Main()'-function arranges the reading and transformation of the requested data from the dataset (steps 1, 2 and 3). After step 3, the csv-dataset is transformed into fragments that contain the coordinates of the requested players (MPOs and static points) for the requested time intervals that need to be compared. These fragments are returned to the 'Main()'-function. In the fourth step, the 'Main()'-function calls upon the 'Compare_fragments()'-function, which arranges the calculation of the QTC-representations of the fragments and the calculation of the distance between those representations. The calculation of the QTC-representations is done by the application of three functions that iteratively calculate the whole matrix sequence of one fragment, one matrix of a sequence and one QTC-relation of a matrix, respectively. These functions only require the coordinates of the fragments, the QTC-variant and the angular tolerance. The 'Compare_fragments()'-function subsequently calls upon the 'Calculate_distance_between_fragments()'-function for each combination of two fragments and stores the returned distances in a distance matrix. It is at this moment that permutations are taken into account, if applicable. The last

[10] The temporal resolution of the analysis is calculated by dividing the temporal resolution of the dataset by the temporal resolution factor. A factor of 1 therefore implicates that the analysis is performed with the original temporal resolution of the dataset.

step of the program consists of the 'Visualise_results()'-function, which takes the distance matrix as argument and returns the requested output (e.g. hierarchical clustering and QTC-matrix sequences) to the user. To avoid redundant information, this paragraph does not mention all of the information of the generalized flowchart shown in Figure 5-1. As mentioned before, the flowchart itself only shows the main processes of the Python program.

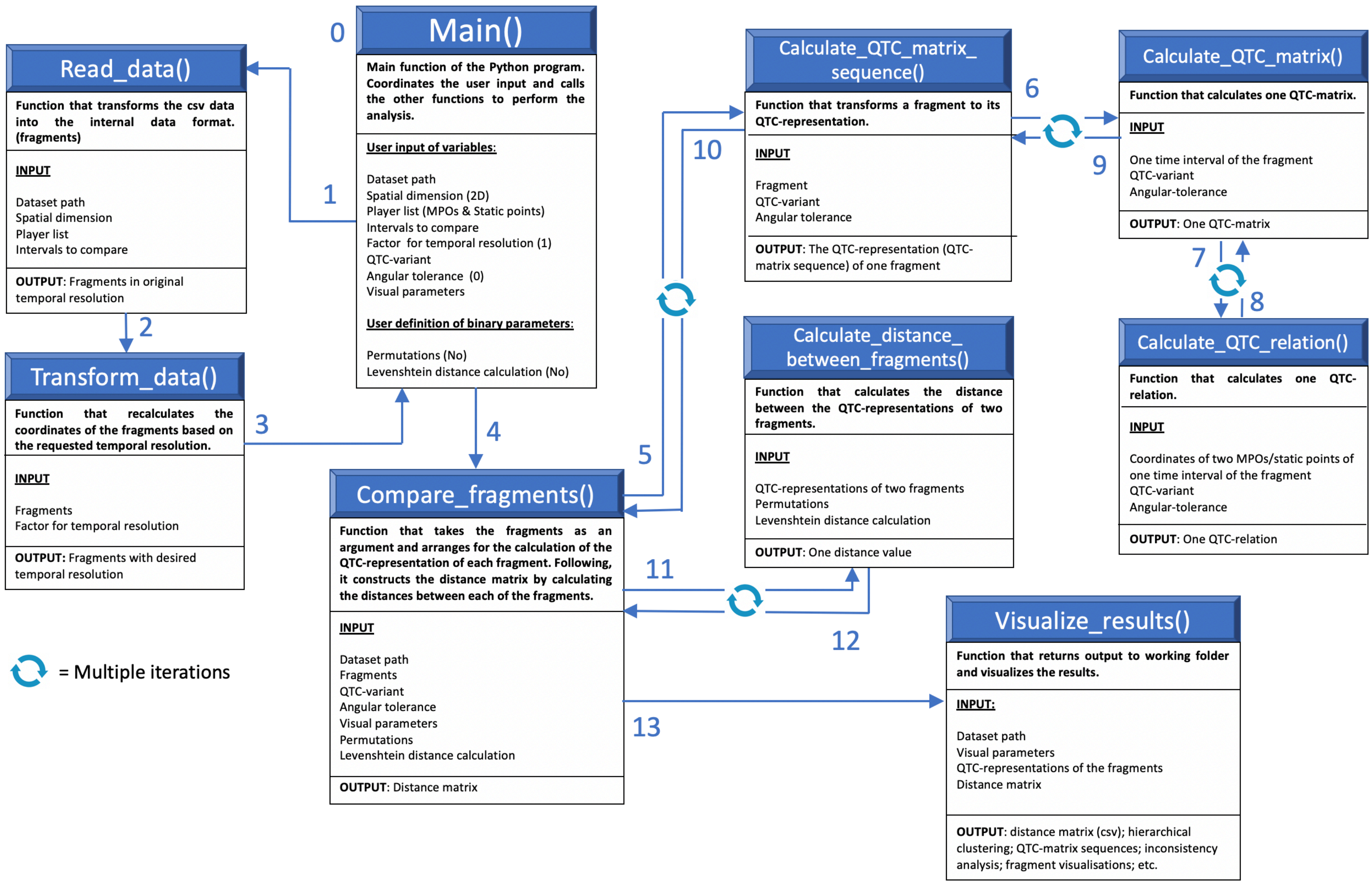


Fig. 5-1. Generalized flow chart of the Python program for dynamic sports analytics.

5.1.2 Static sports analytics

The Python program for the static sports analytics method can be summarized in the generalized flowchart shown in Figure 5-2. This flowchart is constructed in a similar manner as the one for dynamic sports analytics (Figure 5-1), with functions visualized by blue boxes, default parameters in between brackets and the ordering of the functions indicated by light blue numbers. In fact, the processes of both flowcharts show a high resemblance. Performing a static sports analysis starts with the user defining the variables in the 'Main()'-function. Consequently, the 'Main()'-function arranges the reading and transformation of the requested data from the dataset (steps 1 and 2). After step 2, the csv-dataset is transformed into moments containing the coordinates for the requested players (MPOs and, if applicable, static points) for the requested timestamps that need to be compared. These moments are returned to the 'Main()'-function. Note that there is no function or variable for recalculating the temporal resolution of the dataset, since single timestamps are compared with each other. In the third step, the 'Main()'-function calls upon the 'Compare_moments()'-function, which arranges the calculation QTC-representations of the moments and the calculation of the distance between those representations. The calculation of the QTC-representations is done by two functions that iteratively calculate the whole QTC_S-matrix of one moment and one QTC-relation of that matrix, respectively. These functions only require the coordinates of the fragments, the QTC-variant (level one or level two QTC_S) and the angular tolerance.

The 'Compare_fragments'-function hereafter calls upon the 'Calculate_distance_ between_moments'-function for each combination of two moments and stores the returned distances in a distance matrix. It is at this moment that permutations are taken into account, if applicable. The last step of the program consists of the 'Visualise_results()'-function, which is identical to the corresponding function in Figure 5-1.

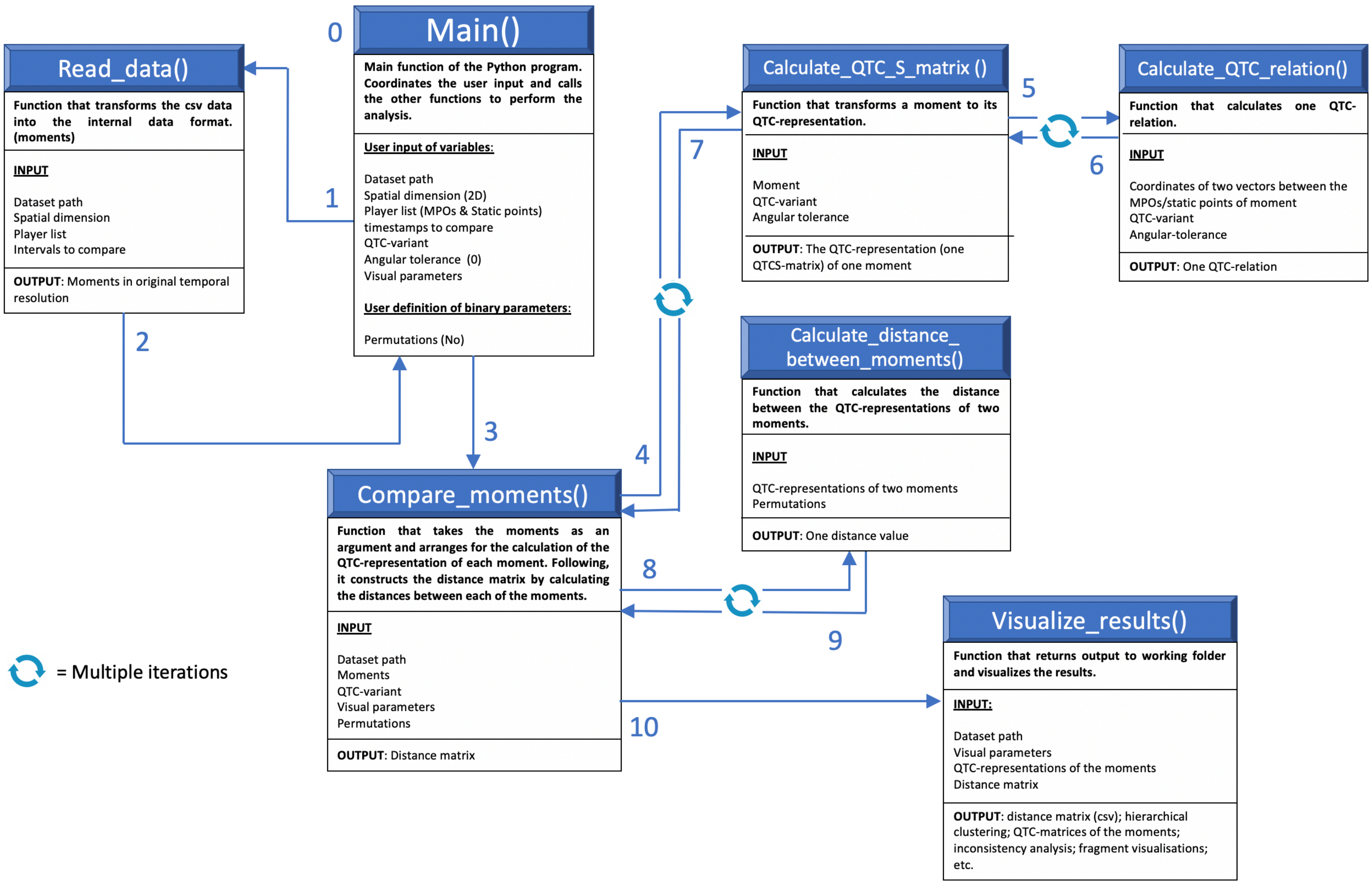
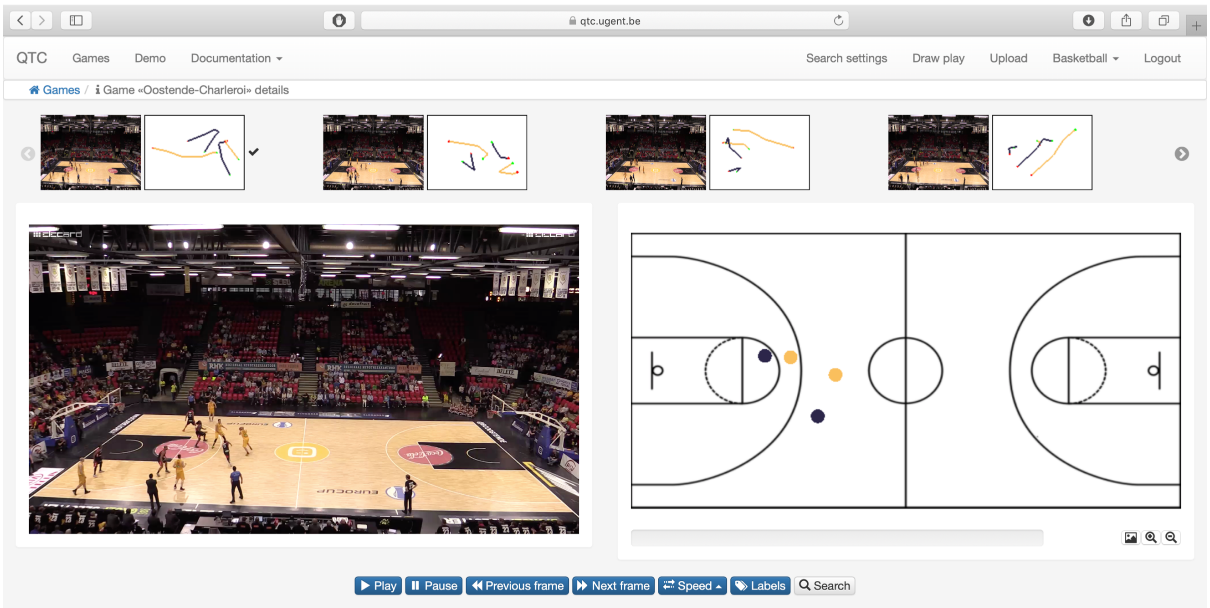


Fig. 5-2. Generalized flow chart of the Python program for static sports analytics.

5.1.3 Front-end

The flowcharts shown in Figures 5-1 and 5-2 present the most important processes of both Python programs. Using these implementations to perform an analysis requires a basic knowledge of the Python programming langue. To enable other users to make queries, a GUI (Graphical User Interface) was designed in cooperation with a third party, which can be accessed through the online web address ‘https://qtc.ugent.be’. Using this online front-end, a user can perform both dynamic as well as static sports analyses, supported by a visual selection of the parameters of the back-end described in Sections 5.1.1 and 5.1.2. A screenshot of the GUI is shown in Figure 5-3, displaying a basketball sports analytics example. A comprehensive manual guide of the GUI is included and can be accessed through the online web address ‘https://qtc.ugent.be/documentation’.



5.1.4 Code complexity

Both programs were written with the aim of enabling experiments to test the use of the dynamic and the static sports analytics methods for analysing real sports datasets. For this reason, little efforts were done to optimize the code and to reduce computing times and storage requirements. Consequently, the code mainly consists of parts programmed using brute force approaches. During the research, however, some thoughts on the optimization of these programs were discovered, mainly aimed at reducing the data volume. As these thoughts are mostly preliminary and do not make up an exhaustive overview of optimization possibilities, they are provided in Appendix 2 rather than here in this chapter.

5.2 Implementation of QTC

Although it is not the goal to discuss all of the functions of the Python programs in detail, it is interesting to discuss the 'Calculate_QTC_relation()'-function to a further extent, given the origin of this research. The 'Calculate_QTC_relation()'-function is present in both Python programs and receives, regardless of its use (for dynamic or static sports analytics), the coordinates of four points as input. It calculates the QTC-relation based on the requested QTC-variant and angular tolerance. In each of the cases, the most basic QTC-variant (level one QTC_B) needs to be calculated, to which additional characters are added in the case of more detailed QTC-variants. There are, however, multiple approaches for calculating this basic QTC_B-relation, given the input of four coordinates. The two approaches that were included in the Python programs are explained in more detail in the next paragraph since, in some specific cases, the choice of the approach can affect the resulting QTC_B-relation for an identical set of coordinates.

To explain both approaches, a theoretical example is used, containing the positions of two MPOs k and l at two timestamps $t1$ and $t2$. To ensure the simplicity of this example, k is chosen to be fixed, so the first character of the QTC_B-relation will always be '0'. Consequently, only the second character of the QTC_B-relation needs to be explained. Further simplifications include an angular tolerance of zero degrees and the assumption of k and l being non-co-located, and with l being non-stationary. The first approach makes use of the Euclidean distance to calculate the QTC_B-relation. To calculate the second character (the QTC_B-relation of *l* with respect to *k*), the Euclidean distance between *k* at *t1* and *l* at *t1* is compared with the Euclidean distance of *k* at *t1* and *l* at *t2*. If the Euclidean distance of the latter is identical to the former, the resulting QTC-character is '0', if it is smaller the resulting QTC_B-character is '-' and if it is bigger it will be '+'. The second approach makes use of the movement direction of the MPOs and calculates, for this example, the movement direction of *l* during the time interval *t1*-*t2* with respect to *k* at *t1*. To do so, the angle (expressed in degrees) between the line connecting *l* at *t1* and *l* at *t2* and the line connecting *l* at *t1* and *k* at *t1* is calculated. If this angle is equal to 90° or 270°, the resulting QTC-character is '0', if it lies in the intervals [0°, 90°[or]270°, 360°] the QTC_B-character is '-' and when the angle lies in the interval]90°, 270°[the resulting QTC-character is '+'. Both approaches are shown in Figure 5-4 where it is demonstrated that, for a specific configuration of begin coordinates and movements, both methods can result in different QTC-relations. For the example presented in this figure, the Euclidean distance approach results in a '+' character, while the angular approach results in a '-' character for level one QTC_B. The reason for this difference is that the displacement of l during the time interval *t1*-*t2* is too big to result in a '-' character for the Euclidean distance approach. While *l* indeed moves towards (*i.e.* in the direction of) *k*, the displacement is so big that the Euclidian distance between *k* and *l* is larger at *t2* then at t1. This issue is similar to the problem discussed in Section 2.3.1 (Figure 2-7), where it was argued that, because of the limited temporal resolution of sports data, the continuity of QTC could not be guaranteed. One could assume that in recent years sports data are recorded with a sufficiently fine temporal resolution, making sure that the problems shown in Figure 5-4 would not occur for the Euclidean distance approach. For the implementation of the dynamic and static sports analytics methods, however, the movement direction approach was chosen as preferred for both Python programs. The main reason for this decision is that, according to us, the assumption that all sports data have a fine enough temporal resolution is rather optimistic. Practice learns that differences in the calculated QTC-characters using both approaches do indeed occur when analysing sports data. For football, for example, players' coordinates are often logged with a temporal resolution of 25Hz, which can be considered as sufficiently fine when compared with the maximum possible movement speeds of the MPOs (*i.e.* football players) on the field, which is reported to be not higher than 10 m/s (Ferro *et al.*, 2014). In the logged datasets, however, coordinates are often duplicated for consecutive timestamps, due to inaccuracies of the measurements (e.g. because of player occlusions, camera vibrations, camera blurring or the limited spatial accuracy of the camera). If such a dataset would be analysed using the dynamical sports analytics method and the original temporal resolution, these identical coordinates for consecutive timestamps would create false ‘0’-characters in QTC_B. Therefore, it is advisable to remove duplicate coordinates or to reduce the temporal resolution of 25Hz to a more coarse one, implying that the situation shown in Figure 5-4 is more likely to occur. Indeed, the same issue is present for coordinate logging systems for other sports, lowering the actual temporal resolution of the recorded datasets. (Bartlett, 2014).

This issue cannot be ignored when applying QTC on sports data. Assuming *k* and *l* are football players, player *l* decides to move in the direction of *k* at *t1* in the provided example. This incentive would, however, not be captured by the Euclidean distance approach. An additional reason for choosing the movement direction approach is that it can be programmed much more efficiently, reducing computing power requirements significantly when analysing large sports datasets.

Consequently, the decision on the approach to calculate the QTC-relations also affects the method for static sports analytics, as the QTC-relations making up the QTC_S-matrices are calculated in a similar manner as their dynamic counterparts (see Section 4.1). Remarkably, the impact of this decision is bigger for static sports analytics, since QTC_S-relations for every pair of vectors between all of the MPOs are calculated. Because of the construction of these vectors, situations as displayed in Figure 5-4 are expected to occur more often. In Figure 4-4, for example, a considerable number of the QTC-characters would be different if they were calculated using the Euclidean distance approach rather than the movement direction approach. The relation between vectors *i* and *c* of moment 1, for example, would be a ‘++’ using the Euclidean distance method, instead of the ‘--’ shown in the QTC_S-matrix of Figure 4-4. Regardless of the chosen approach, it is of utmost importance to only compare QTC-representations that were calculated in the same manner to obtain meaningful results.

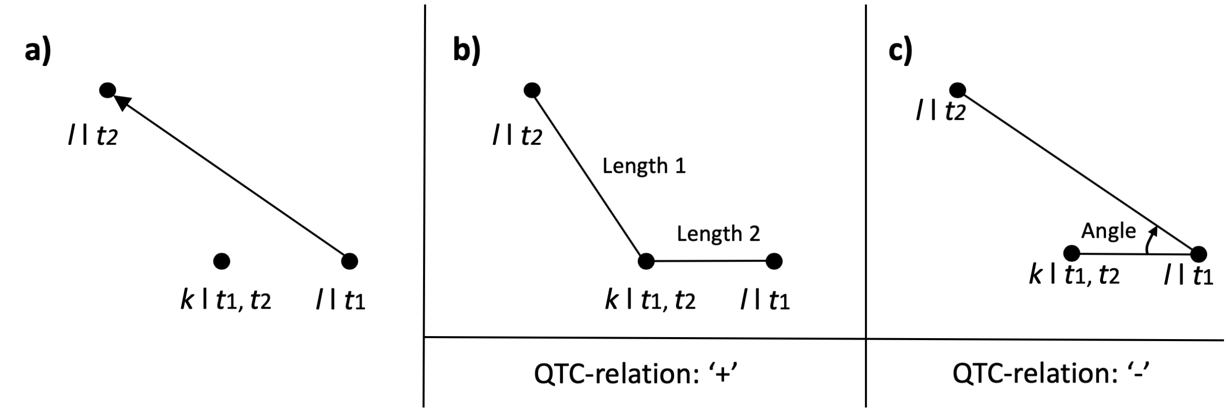


Fig. 5-4. The positions of a (static) MPO k and a non-co-located MPO l at two timestamps t_1 and t_2 (a), the Euclidian distance approach for the calculation of the level one QTC_B-character (b), and the movement angle for the calculation of the level one QTC_B-character (c).

REFERENCES

Bartlett, R. (2014). Introduction to sport biomechanics: analysing human movement patterns. London, U.K.: Routledge.

Ferro, A., Villacieros, J., Floria, P., Graupera, J.L. (2014). Analysis of speed performance in soccer by a playing position and a sports level using a laser system. *Journal of Human Kinetics*, 44, 143-153.

PART II – USE CASES

"A good coach will make his players
see what they can be rather than
what they are"

– Ara Parseghian –

6.

USE CASE 1: DYNAMIC FOOTBALL ANALYTICS

Parts of this chapter were modified from: BEERNAERTS, J., DE BAETS, B., LENOIR, M., VAN DE WEGHE, N. (*under review*). Spatial movement pattern recognition based on relative player movements in soccer . *PLOS one*.

In this Chapter, the use of the proposed method for dynamic sports analytics for team behaviour is investigated, *i.e.* the use of the method for the (automatic) detection of frequent occurring movement patterns between a certain number of players. The aim of this chapter is to provide an answer to research question **R1** and more specifically to research questions **Ra**, **Rb** and **Rc.** For this, a use case is designed, evaluating whether the proposed method for dynamic sports analytics can be used to perform player movement pattern detection in football. In the first section, an introduction to the use case is given. The second section presents an overview of the state-of-the-art for player movement pattern detection in sports, both for team ball sports in general as well as for football more specific. Section 6.3 includes a description of the used dataset and set-up of the use case. The results are presented in the following section, which are validated in Section 6.5. To conclude, Section 6.6 includes a discussion of the results and their value for the domain player movement pattern detection in football. An evaluation of the posed research questions is not included in this chapter, as this is presented comprehensively for all research questions and use cases in Chapter 9.

6.1 Introduction

In football and coaching sciences, the analysis of player movement patterns has primarily focused on the impact of factors such as field size, the number of players (Vilar *et al.*, 2014) on the field and even weather conditions or coach encouragement on the performed movement patterns (Kelly & Drust, 2009; Castellano *et al.*, 2013). At a more individual level, the mental impact of the performed movement patterns was studied intensively (Randers *et al.*, 2017). More recently, methods have been introduced that combine spatial and contextual information to study performance in football (Fernández *et al., 2019*). However interesting, these studies fail to detect actual performed player movement patterns, that can be used by a coach or analyst to describe the playing style of a group of players or of a team. Playing styles in football can be described as the general behaviour of a group of players or of a whole team, that is aimed at achieving the defensive and offensive objectives in a match (Fernandez-Navarro *et al.*, 2016). Impact of different playing styles on team performance can be studied (Tenga *et al.*, 2010; Winter and Pfeiffer, 2010) by analysing one or multiple (Hewitt *et al.*, 2016) metrics that can be derived from positional data such as, amongst others, ball possession (Lago-Ballesteros *et al.*, 2012; Collet, 2013), passing directions and distributions (Tenga and Larsen, 2003) or the locations of events such as interceptions, ball losses and set pieces (Lago-Peñas *et al.*, 2017). However, performed playing styles in football are influenced by a variety of factors, where even contextual factors such as match status or match venue have proven to be of importance (Fernandez-Navarro *et al.*, 2018). Besides the study on playing styles, neural networks and machine learning techniques have been applied on football to detect patterns of tactics based on positional data (Leser *et al.*, 2018; Perl *et al.*, 2013). Detection of reoccurring player movement patterns could contribute to the characterization of playing styles and tactics in football by describing the movement behaviour of players on the field. In football clubs, player movement pattern detection is generally done by video inspection and notational analysis (James, 2006) of both the own team as well as the opponent. The main aspect of this detection consists of visually finding and annotating similar movement patterns that occur during one or more games, which is a very time-consuming and subjective effort that often has a limited quality due to low observational accuracies of coaches or analysts (Nicholls & Worsfold, 2016). Methods to detect player movement patterns through the analysis of match events in football have been developed (Sarmento *et al.*, 2010; Sarmento *et al.*, 2014). However, to the best of our knowledge, and as argued by Feuerhake (2016), an optimal player movement pattern detection method aimed at the automatic detection of reoccurring spatial movement patterns of soccer players, is currently not available.

When analysing the movements of football players during a football match, it is of interest to detect movement patterns of one or more players. This means finding different time intervals, in one or multiple games, during which one or more players performed similar movements on the field. This is possible with or without predefining the movement pattern of interest. When no such pattern is used as reference, the analysis can be seen as an example of data mining (Feuerhake, 2016; Jain *et al.*, 2000). Concretely, it implies that similarities between all possible movements that occur during one or multiple football matches are calculated. Possible results include one or multiple groups of movements that have similarities higher than a certain threshold. Although data mining approaches were used to detect movement patterns in football (Grunz *et al.*, 2012; Gudmundson and Wolle, 2014; Niu *et al.*, 2012), they do not guarantee that the detected movement patterns are meaningful for soccer coaches, analysts or reporters (Zhang *et al.*, 2016; Memmert & Raabe, 2018). When a predefined spatial movement pattern is used as reference, the problem is referred to as pattern matching. Applied on football, pattern matching implies that similarities between the reference movement and all other movements that occur during one or multiple soccer matches are calculated. Consequently, a player movement pattern is detected when a time interval with a distance lower than a given threshold is found.

This case study is designed according to this pattern matching principle, following the terminology that was introduced in Section 3.1. This means that the player movements during the predefined time interval are named as the ‘reference fragment’. The aim of applying the dynamic sports analytics method is to support a football coach or analyst to select a specific game fragment as reference fragment and search a database (containing a limited number of other fragments) for similar movement patterns for the same or other players. Accordingly, a player movement pattern is detected when a fragment with a distance lower than a given threshold with respect to the reference fragment is found. The main reason for this approach is to allow the user to decide for him/herself what the pattern of interest is, ensuring that the results have an added value from the sports perspective (Memmert & Raabe, 2018).

6.2 Player movement pattern detection in sports

Research on movement pattern detection and performance analysis in sports has seen a huge rise of interest in recent years, as sports data became more widely available by virtue of technological developments, decreasing prices of tracking technologies and sports federations adopting policies for application in official matches (D'Orazio & Leo, 2010; Link, 2018). Match analysis in soccer originally relied on video images and consisted of quantitative assessments (e.g. pass frequencies) and qualitative assessments such as experts evaluations (Memmert & Raabe, 2018). With the current state-of-the-art tracking technology it is possible to log the (x,y,t)-coordinates of football players by means of cameras that are mounted around the field (Rein & Memmert, 2016), thus preventing any influence on the actual performed movements because no physical trackers are placed on the players (Figueroa *et al.*, 2006). Tracking of the ball, however, remains a challenging task in different sports, because of the high speed of the ball and the regular occurrence of occlusion (Kelley *et al.*, 2010; Castellano *et al.*, 2014; Gomez *et al.*, 2014). For this reason, the ball is not taken into account in the case studies presented in this dissertation. As described by Memmert and Raabe (2018), positional data of soccer players allows for physiological and technical assessments (match analysis level 3.0) and dynamical and tactical assessments (match analysis level 4.0) of football matches.

With the trajectories of the players available during the match, the question arises how to analyse the player movement patterns (match analysis level 4.0), in order to enhance the players' and team's performance and, ultimately, win more games. Since player movement pattern detection in various sports uses the same type of data, we first briefly present the state-of-the-art in team ball sports in general before presenting a more focused review of the established methods in soccer.

6.2.1 Team ball sports

Jonsson *et al.* (2010) introduced a method (T-pattern) that combines positional and event data to find temporal patterns in various sports. Preliminary examples show that this method could be used for pattern detection in team ball sports such basketball and football. However, the method aims at finding patterns for one player at a time, thereby not taking relations between different players into account. In beach volleyball, Pérez-Turpin *et al.* (2009) and Seweryniak *et al.* (2013) reported about their search for a methodology for detecting movement patterns. For beach volleyball, however, movement pattern recognition is limited, as each team consists only of two players. Player movement patterns of multiple (especially more than 2) players have been detected and analysed based on the trajectories of players in basketball. Space-time movement patterns of both playing dyads as well as whole teams in basketball (using a stretch index based on the geometric centre of each team), were studied by Bourbousson *et al.* (2010; 2010b). Focusing on one aspect of the game, Leite *et al.* (2014) studied the effects of defensive pressure on the performed player spatial movement patterns. Sha *et al.* (2016) introduced the principle of 'Chalkboarding' in sports analytics, where a coach or analyst can draw the requested pattern (called 'a play') and the system returns the time intervals during which the players performed similar movements. Sha *et al.* use the Euclidean distance between the positions of corresponding players at each timestamp of different plays to calculate the distance between the different plays. As such, identical plays performed on different parts of the field have a relatively high distance from the reference play. Sweeting *et al.* (2017) proposed a method for detecting player movement patterns in netball based on a sequence analysis of secondary parameters (derived from the players' trajectories) used for describing the external load on players.

6.2.2 Football

Football has attracted substantial research interest in the past years, by academics as well as by private companies. In this chapter, however, we focus on a rather small subcategory of sports analytics in football, *i.e.* player movement pattern detection. For a more general overview of sports analytics in football, we refer to the works of Rein & Memmert (2016), Sarmento *et al.* (2014) and Memmert *et al.* (2017).

Most research in player movement pattern detection in football makes use of quantitative approaches. Kang *et al.* (2016), for example, implemented a method for evaluating the strategic performance of players, based on the regions of the field players consecutively move through during the game. One of the challenges is the division of a football match into fragments that are of interest to the analyst. Trajectory segmentation, for example, is a research domain concerned with finding objective criteria to split trajectories of moving objects during long periods (*i.e.* football players during one whole football match) into smaller fragments (Buchin *et al*, 2011; Buchin *et al.*, 2013). Similar to the team centroid method of Bourbousson *et al.* (2010; 2010b) in basketball, efforts have been made on studying spatial movement patterns of a whole team (Lames *et al.*, 2010; Duarte *et al.*, 2013), player groups (Gonçalves *et al.*, 2014) or individual players (Sampaio and Macãs, 2012) using the team centroid in football. An interface that supports a football analyst with identifying interesting game situations in football was created by Shao *et al.* (2016). Other efforts for facilitating the visual abstraction of player movement patterns, using clustering methods such as *k*-Means and *k*-Medoids, can be found in Sasha *et al.* (2017). Popular quantitative distance measures for player movement pattern detection in football are, amongst others, the average Euclidean distance (Nanni & Pedreschi, 2006), the perpendicular and angle distances (Lee *et al.*, 2007) or the Fréchet distance (Fréchet, 1960). Gudmundsson & Wolle (2014), for example, use these distance measures for detecting correlations between player movements.

A smaller number of studies are qualitative, with considerable efforts on the detection of player spatial movement patterns through the use of T-patterns (Magnusson, 1996;

2000). The T-pattern method aims at revealing temporal patterns that are not detectable through classic visualisation techniques. It was applied on soccer by, amongst others, Sarmento *et al.* (2010) and Camerino *et al.* (2012) to detect temporal patterns of match events. Feuerhake (2016) and Feuerhake & Sester (2013) try to discover unknown player movement patterns in football data by using data mining techniques. At first this was done only for multiple players (Feuerhake & Sester, 2013) but later the method was generalized and extended to the trajectories of individual players (Feuerhake, 2016). The biggest difference of this approach with our proposed method for dynamic sports analytics is that in the former, the positions of the players at single timestamps are placed in sequence rather than the movement/displacement during the time intervals between the timestamps. A similar approach was used earlier by Grunz *et al.* (2012), who used static team formations and their temporal evolution for player movement pattern detection. Feuerhake later suggested, however, to also use the movement during the time intervals as sequence elements as is done in our novel method, be it only for single players. Also, using the approach of qualitative description of the movements of multiple players during time intervals, Relative Movement (REMO), introduced by Laube *et al.* (2004), can be applied for movement pattern detection. Laube *et al.* included a football example in their paper, but to the best of our knowledge, neither this nor the other methods mentioned above were adopted by coaches or analysts of professional football teams, nor were they proven sufficiently effective for player movement pattern detection in football (Feuerhake, 2016).

6.3 Dataset and analysis set-up

In this section, the dataset and of the set-up of the experiment are discussed, giving a detailed overview of the use case.

6.3.1 Dataset

The dataset of this use case study stems from a real football match of a 2016-2017 professional football competition. Due to privacy concerns, the teams and players are presented anonymously. During the match, players were tracked by means of a camera system that was mounted in the stadium, logging their (x,y,t)-coordinates with a temporal resolution of 25Hz. The resulting dataset consists of 144,086 (x,y,t)-coordinates for each of the 22 players on the field. Besides this, a manual annotation of the events occurring at each of the timestamps of the match is available (XML-file). Using the event information, positional data is filtered, making sure that all considered movements (fragments) fully occurred during the actual game (e.g. not when the ball was out of play or when an injury treatment was taking place).

6.3.2 Analysis set-up

By considering football players as MPOs, QTC can be used to describe the movements of those players during a particular time interval in a match. Following the terminology introduced Sections 3.1 and 6.1, each fragment can be described by a QTC-matrix sequence where the number of players determines the dimension of the individual QTC-matrices that make up the sequence, while the sequence length is defined by the temporal length and resolution of the fragment. In this use case, a movement pattern recognition example, there are two types of fragments. First, there is the reference fragment, which is the fragment of interest to the coach or analyst, *i.e.* for which (s)he wants to find similar fragments in the database. The second type of fragments are the target fragments, which are all other fragments of interest, *e.g.* all fragments for the same players. To start with, the reference fragment is transformed into its QTC-matrix sequence representation. Subsequently, all target fragments are also transformed to their QTC-matrix sequence representations. After that, the distances between the reference fragment and all target fragments are calculated by comparing the QTC-matrix sequences. Target fragments with small distances to the reference fragment are considered to be more similar to the reference fragment than target fragments with larger distances. The result of the player movement pattern detection will thus be a list of the target fragments ordered according to the calculated distance. A coach or analyst can use this list to examine the fragments that contain the movements most similar to the reference fragment (s)he has selected, and can discover whether there is a pattern that occurs regularly either by his/her own team or the opponent team. Ultimately, (s)he can then adjust coaching to increase the team performance and win more matches.

Given the computational limitations (Section 5.1.4) we chose a situation where a coach or analyst selects a rather simple reference fragment consisting of two players. A simple reference fragment is chosen to ensure that relatively similar movements might be found in just one soccer match (Feuerhake, 2016). Straight sprinting towards the opponents goal area, for example, is a regularly trained player spatial movement pattern (Jeffreys, 2008) and is in fact the most common pattern to occur before a goal for both the scoring and the assisting player (Haugen *et al.*, 2013). For this reason, the first occurrence of such a straight run of 2 players in the match was selected as reference fragment. The 20 seconds (500 timestamps) following the start of the sprint, a common duration for an attacking action in soccer (Rafael *et al.*, 2017), were included in the reference fragment to add some movement complexity besides the straight run. The reference fragment (Figure 6-1) thus contains the movements of two attacking players of the same team, and starts with them quickly moving towards the opponent's goal area. Around the sixth second of the fragment, however, the ball possession is lost and the two players start a slower defensive (parallel) run towards their own goal area. A relatively simple approach is chosen for the trajectory segmentation of the football match of interest (Buchin *et al*, 2011; Buchin *et al.*, 2013). For the full length of the match, every 10 timestamps (0.4 seconds) a target fragment of 500 timestamps is created, resulting in a set of 14,408 target fragments. Each target fragment contains the movements of the same players as the reference fragment, during 20 seconds. The temporal resolution of both the reference fragment as well as the target fragments is reduced with a factor 10 to facilitate faster calculations and to reduce the impact of noise in the data on the pattern detection (Feuerhake, 2016). To ensure a correct rotation of both the reference as well as the target fragments (*i.e.* going towards the opponent's goal area in the beginning of the fragment), 4 static points are included into the QTC-description of all fragments. The number and locations of the static points are chosen according to the guidelines mentioned in Section 3.3. The 4 static points are placed on the corners of the football field, as is shown for the reference fragment in Figure 6-1.

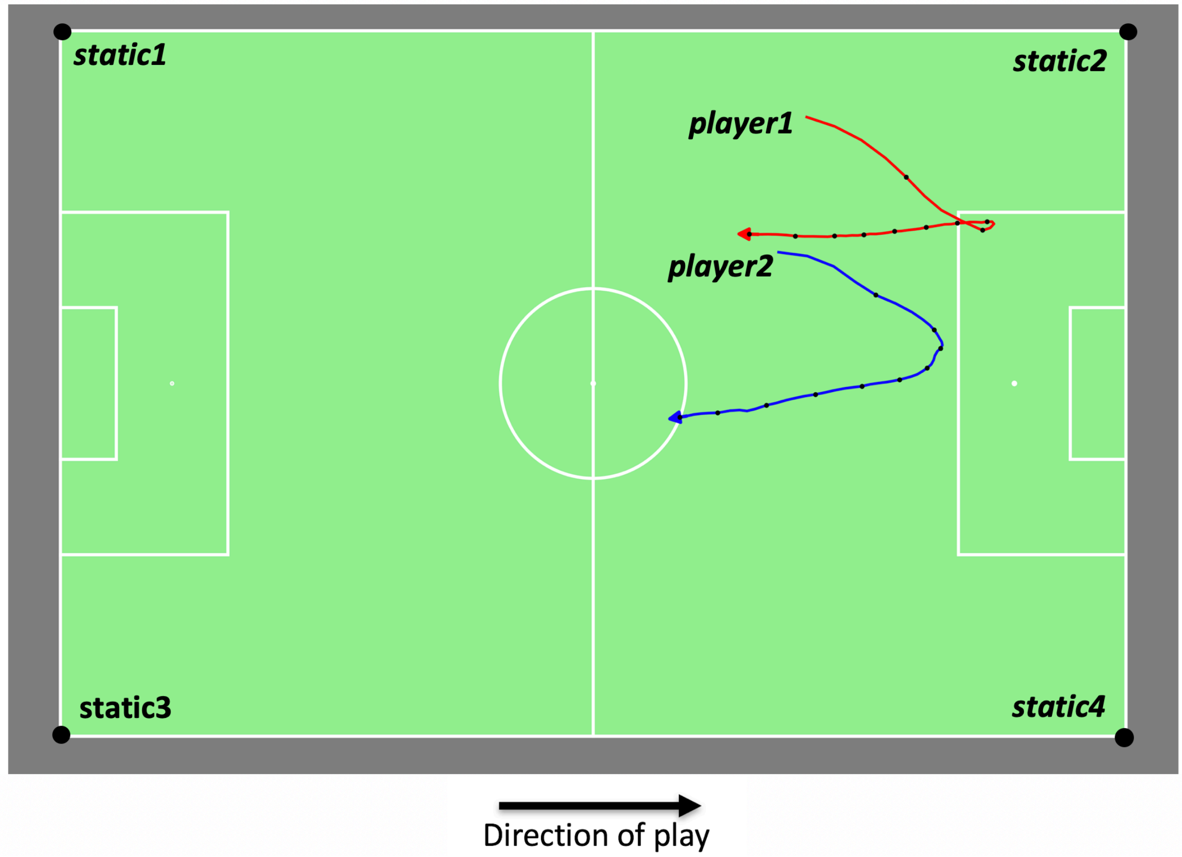
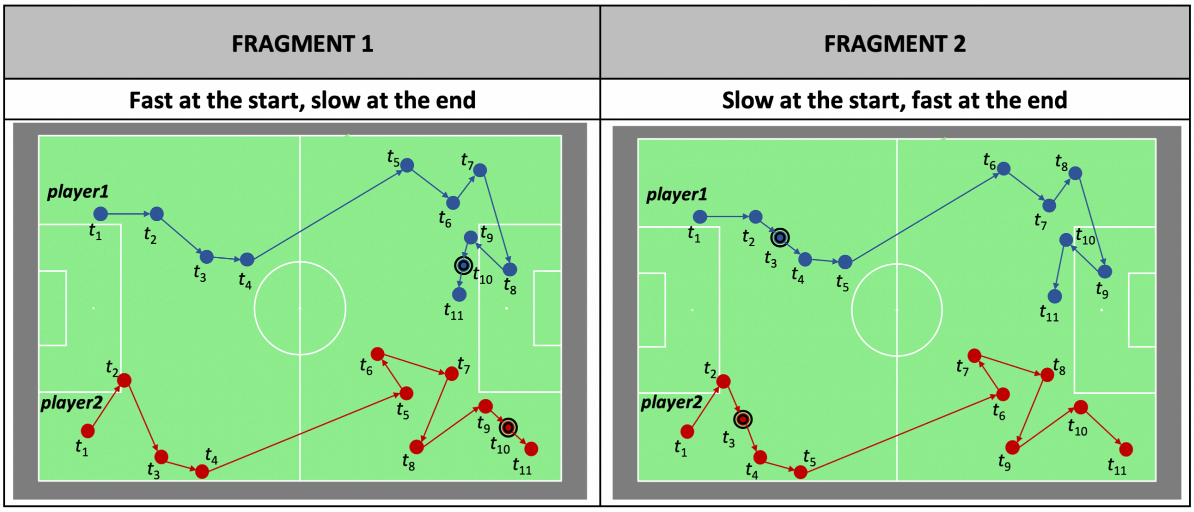


Fig. 6-1. Reference football fragment of 20 seconds. This fragment contains the movements of two players (*player1* and *player2*) together with four static points (*static1*, *static2*, *static3* and *static4*). Small black dots on the trajectories indicate the positions of the players at every 2 seconds.

When using only two MPOs (*Player1* and *Player2*), the more detailed QTC_C-variant would be advised over QTC_B for the proposed movement pattern detection method, to ensure results are meaningful (see Section 3.4 and Figure 3-7). However, because 4 static points are included in the QTC-description of the fragments, the more basic QTC_{B21}-variant is chosen. The motivation for this is that this limits the number of QTC-characters in the QTC-descriptions of the fragments, consequently reducing both the risk of over-averaging (see Section 3.3) as well as computing times. Furthermore, no angular tolerance for the 0-character is used ($\alpha = 0°$), given the limitations mentioned in Section 3.5.

It can be assumed that the football coach or analyst demands an exact match of the players for this use case, meaning *player1* and *player2* should not switch roles in the target fragments. Because of this assumption, permutations are not included in the distance calculation between fragments. If permutations would be allowed, however, target fragments having similar movements with the reference fragment, but with *player2* playing more on the flank and *player1* playing more in the centre of the football field, would have low distances to the reference fragment.

Finally, it is important to choose whether or not to apply the Levenshtein distance metric. Because reference and target fragments have identical lengths, it is possible to calculate distances without the use of this method. However, given the changes in pace of the movements that make up the reference fragment (see Figure 6-1), this is not advised. If two fragments contain identical movements, performed at different speeds (e.g. a higher speed at the beginning and a lower speed at the end, as for fragment 1 in Figure 6-2), the Levenshtein distance metric will give an appropriate penalty in the distance calculation. As such, the calculated distance between these fragments will not be zero, but will be smaller than the distance that would result from a distance calculation based on pairwise comparisons (*i.e.* simple similarity, where no substitutions or insertions of QTC-matrices are allowed). Figure 6-2 shows two of such fragments along with the distance between them (using QTC_{B21}), computed with both the Levenshtein distance metric as well as with the pairwise-comparison distance calculation. It illustrates the ability of the first method to produce a more suitable, in this specific case lower, distance. Consequently, when two target fragments contain movements identical to the movements of the reference fragment, but with movement in one target fragment occurring at a different pace, the Levenshtein distance metric will produce a higher distance for the latter. The Levenshtein distance metric therefore takes differences in pace of movements in account when comparing fragments, which results in more suitable distance values, as was argued in Section 3.8. For this reason, we choose to use the Levenshtein distance metric in this use case.



Distance between both fragments based on pairwise comparisons	**0.5**
Distance between both fragments based on Levenshtein	**0.2**

Fig. 6-2. Two almost identical fragments with movements of two players (*player*1 and *player*2) with small differences in speed along with the distances calculated between them based on pairwise comparisons and on the Levenshtein distance metric (using QTC_{B21}). The difference in speed occurs at t_{10} and t_3 in fragment 1 and 2 respectively, and is accentuated by a double circle around the players' positions.

6.4 Results

Since for each of the target fragments the distance with respect to the reference fragment is calculated, they can be ranked according to this distance. Fragments with a low distance contain movements that are quite similar to the movements in the reference fragment, while the movements in fragments with a high distance will hold little resemblance to the reference fragment. Figure 6-3 displays the nine most similar, non-overlapping, target fragments ordered according to ascending distance. Non-overlapping fragments are obtained by ordering the fragments according to their rank (starting from rank number 1) and omitting the fragments with higher ranks from the final ranking when they overlapped with a fragment with a lower rank. Also, the resulting fragments are filtered to check whether they fully occurred during the game (thus not during breaks, see Section 6.3.1). As such, the number of results is reduced to 221. Depending on the actual movements in the match, a top-*k* (the *k* top ranked target fragments) can be considered as highly similar to the reference fragment (similar to the approach of Sha *et al.* (2016)). To evaluate the results of this use case and their significance, however, objective validation is required and will be presented in the next section.

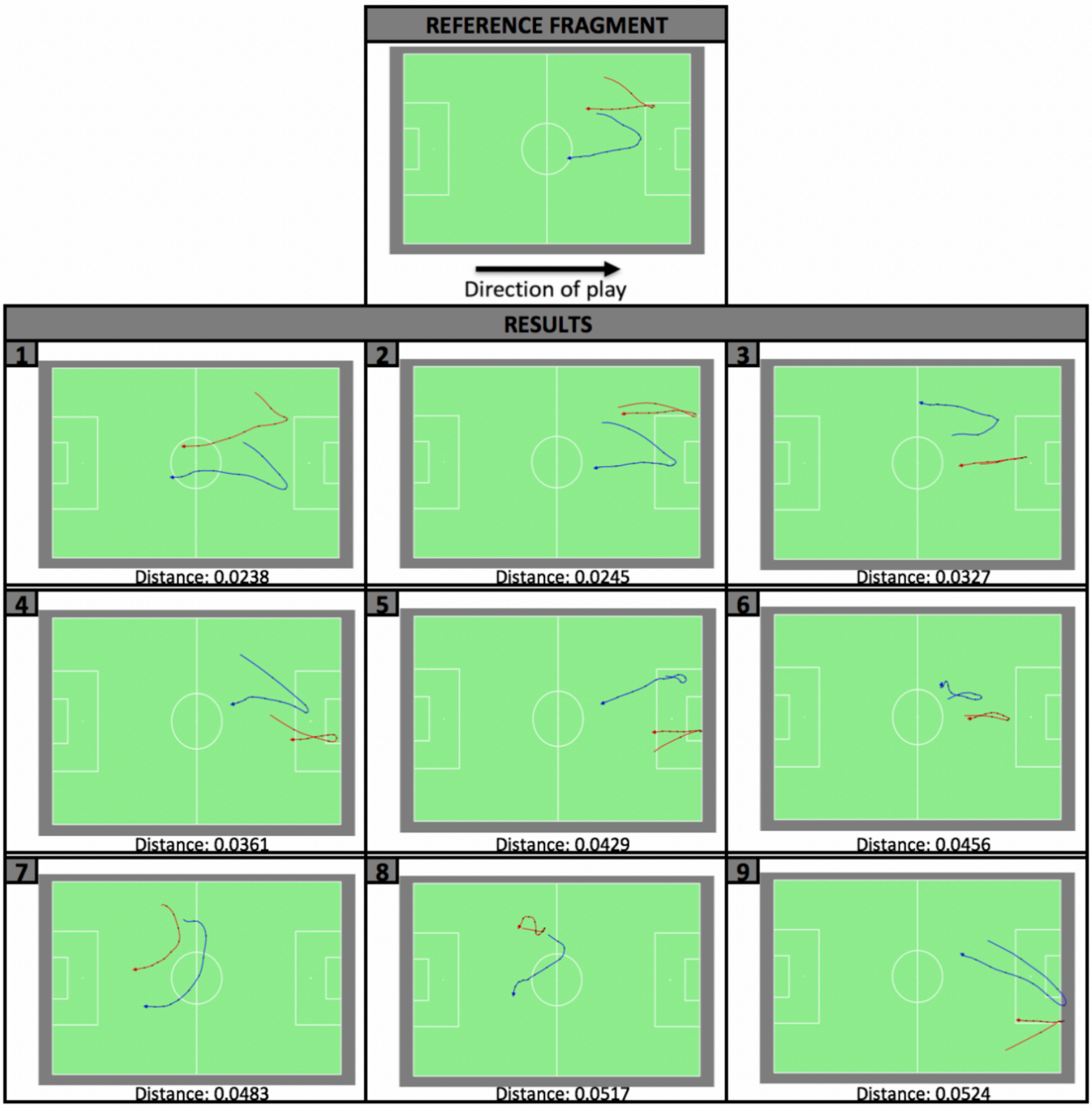
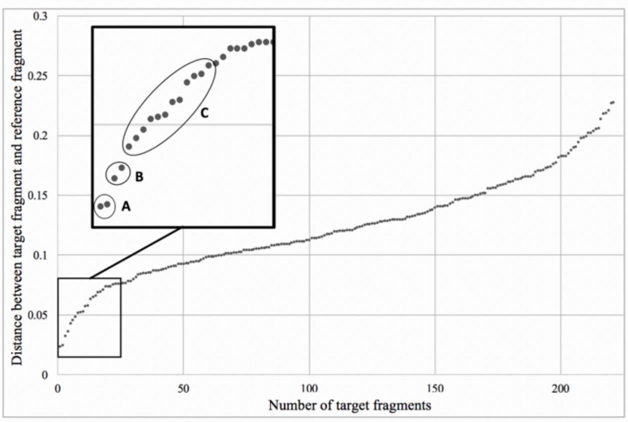


Fig. 6-3. Top nine results of the player movement pattern recognition with the reference fragment shown on top. The results are ordered according to ascending distance to the reference fragment, with their rank numbers noted on the left top of the individual visualisations.

6.5 Validation

As Feuerhake (2016) argues, the verification of movement pattern detection methods and their results in sports is not a straightforward task, due to the lack of a good ground truth. Furthermore, due to the huge variety in different methods, it cannot be assumed that finding the same results as other established methods is desirable nor that it should be the goal. As such, conform to other studies in this field, validation was firstly done by visual comparison of the top-*k* results with the reference fragment. After successfully validating the results in this manner, we proceeded by statistically testing the ranking of the top-*k* results presented in Figure 6-3. To that end, we ordered the non-overlapping target fragments according to their distance with the reference fragment and could detect two distinct groups of fragments (groups A and B in Figure 6-4a), containing the fragments with rank numbers 1 & 2 and 3 & 4, respectively. Considering the regularly increasing distance curve for fragments with rank numbers above 4, a control group C was created containing 12 elements (see Figure 6-4a). For validation, we wanted to test whether a panel of football experts would confirm the ranking in Figure 6-3. The expert panel consisted of 37 bachelor students of the Department of Movement and Sports Sciences of Ghent University with a good knowledge of football. In their curriculum they have had at least two years of football classes, including practical sessions and theory on technique and tactics of football. The study, which was conducted in the spring of 2018, received institutional approval and the participants' informed consent was obtained. As the test in fact only consists of a visual assessment of spatial patterns, without taking implications for coaching or performance enhancement in soccer into account, the bachelor students can indeed be seen as an expert panel for the given test. As a statistical test, a duo-trio test (Peryam & Swartz, 1950; Meilgaard *et al.*, 2007) was set up, a statistical test for determining whether a difference exists between two samples, by asking which of the two samples most resemblances a reference fragment. The duo-trio test consisted of 18 questions (Figure 6-4b) which were presented to the participants in random order. Each question contained the visualisations (as in Figure 6-3) of the reference fragment and two sample fragments, of which the participants had to indicate which one was most similar to the reference fragment. Six questions compared the fragments of groups A and B (1 question A vs A, 1 question B vs B and 4 questions A vs B), twelve questions compared the union A∪B with group C. In the latter series of questions, each fragment of C was randomly combined with a fragment of A∪B. Questions for which the participants chose the sample fragment with the lowest rank number as most similar to the reference fragment were considered as correct answers. The information and questionnaire that was presented to the members of the expert panel together with an overview of the answers for each question for each of the members of the expert panel are included (anonymously) in Appendix 3 of this dissertation.

a)



b)

	Rank number fragment 1	Rank number fragment 2	Number of participants	Number of correct answers	Level of significance
A∪B vs C	3	6	37	26	1.0%
	10	3	37	28	1.0%
	8	4	37	32	0.1%
	1	9	37	35	0.1%
	2	13	37	37	0.1%
	16	2	37	34	0.1%
	4	12	37	32	0.1%
	5	1	37	34	0.1%
	4	15	37	32	0.1%
	11	3	37	26	1.0%
	2	7	37	35	0.1%
	14	1	37	33	0.1%
A vs B	3	2	37	36	0.1%
	3	1	37	30	0.1%
	2	4	37	31	0.1%
	1	4	37	25	1.0%
B vs B	3	4	37	10	/
A vs A	2	1	37	12	/

Fig. 6-4. Set-up and results of the duo-trio test. The non-overlapping fragments ordered according to their distance with respect to the reference fragment, and the delineation of the groups A, B and C (a). The questions, answers and significance levels of the duo-trio test (b).

The results of the duo-trio test, aggregated per question, are presented in Figure 6-4b. The results show that the participants validate the ranking positively, thereby distinguishing between the fragments of groups A∪B and C and between the fragments of groups A and B with high levels of significance (Stone *et al.*, 2012). When it comes to ordering the elements in groups A and B separately (A vs A and B vs B), the participants do not validate the exact ordering presented in Figure 6-3. Considering the small differences in distance between the fragments in those groups, especially compared to the distances with the other fragments, this can be considered as acceptable. As such, the top-*k* (with *k*=4) is validated by the duo-trio test (Stone *et al.*, 2012).

6.6 Discussion

In this section, the importance of the results for the domain of team behaviour analysis, player movement pattern detection in football more specifically, is discussed. An evaluation of the use case for the sports analytics domain and its value for answering the respective research questions can be found in Chapter 9.

In literature, it is described that, because of the limited accuracy of tracking technologies and the vast number of possible positions of players on the football field, it is almost impossible to find two fragments that contain identical movements (Feuerhake, 2016). For this reason, a rather simple reference fragment is required in order to find at least one relatively similar target fragment in the same football match. If such a target fragment cannot be found, it is impossible to validate the use of the proposed sports analytics method for player movement pattern detection in football. Despite its simplicity, however, the reference fragment of the use case can be considered as meaningful and consists of an interesting situation for a football coach or analyst (Jeffreys, 2008; Haugen *et al.*, 2013). At least two highly similar target fragments are found, as the results with rank numbers 1 & 2 (group A) consist of movements that indeed look very similar to the movements that make up the reference fragment. As such, regardless of its rather simple set-up, the use case can be seen as an essential starting point to illustrate the proposed method for player movement pattern detection in football, before switching to more complex examples with more players and more complex movements in the future.

With respect to the limited accuracy of tracking technologies used in football, it can be noted that QTC is a relative calculus, meaning that only the relative movements in two fragments need to be the same in order for them to be considered as identical. Small differences in the players' coordinates of a given movement pattern, for example, do not cause large differences its QTC-matrix sequence representation. Furthermore, QTC allows for the detection of movement patterns consisting of identical movements that occur on different places on the field or at different spatial scales. Adding to the qualitative nature of the calculus, both the temporal (as in the case study) as well as the spatial resolution can be reduced to cope with the limited accuracy of football data. Looking at the results of the use case (Figure 6-3), it can indeed be seen that it is not straightforward to produce a long list of highly similar fragments of the same match using the proposed method. However, when equipped with a big dataset (containing multiple football matches or seasons and thus a significantly higher number of target fragments and a clear focus or definition of the pattern of interest, more specific results, and thus better insights for the football coach or analyst, are to be expected. When it comes to validation of the results, the validation approach consisting of a duo-trio test with a football expert panel and statistical validation can be considered as both trustworthy (sufficient) as well as innovative in the domain of sports analytics, when compared to the validation methods used in similar research (see Section 6.2). In similar research, results are often presented by themselves, thus without any form of statistical validation.

From a methodological perspective, differences in results with similar methods are to be expected. In contrast with the work of Shao *et al.* (2016), for example, the proposed method is characterized by its scale, rotation and translation invariance. This allows to find highly similar movement patterns that occur, for example, on different parts of the football field or at a different scale. Different static points could be included, however, to fix orientation and scale. Furthermore, while QTC typically describes relations between two or more MPOs, the incorporation of static points allows for the comparison of trajectories of single football players (Jonsson *et al.*, 2010; Sarmento *et al.*, 2014; Feuerhake, 2016). Besides this, permutations between players can be included, allowing players to switch roles in target fragments. Furthermore, the Levenshtein distance metric supports the comparison of fragments with different temporal lengths. This means that the target fragments can have a different length than the reference fragment, something that was not included in the case study for reasons of simplicity, and could raise the chance of finding target fragments with lower distances with respect to the reference fragment and thus more similar movement patterns. An important advantage of the proposed method for player movement pattern detection in football is therefore that it allows for the detection of similar player spatial movement patterns that have different temporal lengths or that occur at different speeds. Moreover, we believe that the proposed method grasps the essence of the way a football player positions himself on the field during a game. To that end, (s)he will mostly look at his teammates and opponents, assessing their relative positions to him and adjust his location where needed by moving around on the pitch. This is exactly what QTC describes, *i.e.* the changes in relative positions between all of the players during the match. Most other studies, however, such as Sweeting *et al.* (2017), only use movements with respect to the field (e.g. ‘turn 45 degrees on the field’) for the movement pattern detection, and do not take these interactions between players into account. Limitations of the proposed method for player movement pattern detection in football include the high requirements of computation power and the necessity of coordinates that are recorded with a fine temporal resolution. Another limitation is that ball possession is not taken into account, which could be solved by adding a postprocessing step.

To conclude this chapter, we are convinced that the proposed method for dynamic sports analytics could support a football coach or analyst to conduct an analysis on team behaviour and thus to get insight into the movement patterns that occur within one or multiple football matches. With the proposed method, a coach can easily search for a movement pattern, to check whether the patterns that were trained were executed during official matches, or get insight in patterns played by opponent teams. Future research should be performed on bigger datasets, with more players and more complex movement patterns, with a more rigorous evaluation of the results by real coaches. Besides a validation of the results, a validation of the usability of the detected patterns should be performed. Furthermore, expanding the approach to the field of data mining, it would be interesting to compare all fragments with each other and thus get insights in frequently played patterns by certain players. These results could then be matched with formations of other teams, e.g. which players perform specific movement patterns when playing against other specific players. To achieve such results with reasonable computing times, it will be necessary to implement techniques for code optimization, as suggested in Section 5.1.4

REFERENCES

Bourbousson, J., Sève, C., McGarry, T. (2010a). Space-time coordination dynamics in basketball: Part 1. Intra-and inter-couplings among player dyads. *Journal of Sports Sciences*, 28(3), 339-347.

Bourbousson, J., Sève, C., McGarry, T. (2010b). Space-time coordination dynamics in basketball: Part 2. The interaction between two teams. *Journal of Sports Sciences*, 28(3), 349-358.

Buchin, M., Driemel, A., Van Kreveld, M., Sacristán Adinolfi, V. (2011). Segmenting trajectories : a framework and algorithms using spatiotemporal criteria. *Journal of Spatial Information Science*, 3, 33-63.

Buchin, M., Kruckenberg, H., Kölzsch, A. (2013). Segmenting trajectories by movement states. In Proceedings of the 22nd ACM SIGSPATIAL International Conference on advances in Geographic Information Systems (pp. 351-360). New York, NY, U.S.A.: ACM

Camerino, O., Chaverri, J., Anguera, M.T., Jonsson, G.K. (2012). Dynamics in the game of soccer: Detection of T-patterns. *European Journal of Sport Science*, 12(3), 216-224.

Castellano, J., Alvarez-Pastor, D., Bradley, P. (2014). Evaluation of research using computerised tracking systems (amisco and prozone) to analyse physical performance in elite soccer: a systematic review. *Sports Medicine*, 44, 701-712.

Castellano, J., Casamichana, D., Dellal, A. (2013). Influence of game format and number of players on heart rate responses and physical demands in small-sided soccer games. *The Journal of Strength and Conditioning Research*, 27(1), 1295-1303.

Collet, C. (2013). The possession game? A comparative study of ball retention and team success in European and international football. *Journal of Sports Sciences*, 31(2), 123-136.

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

D'Orazio, T., Leo, M. (2010). A review of vision-based systems for soccer video analysis. *Pattern Recognition,* 43(8) 2911-2926.

Fernández, J., Bornn, L., Cervone, D. (2019). Decomposing the Immeasurable Sport: A deep learning expected possession value framework for soccer. In Proceedings of the 13th annual Sloan Analytics Conference. Boston, MA, U.S.A.

Fernandez-Navarro, J., Fradua, L., Zubillaga, A., Ford, P.R., McRobert, A.P. (2016). Attacking and defensive styles of play in soccer: analysis of Spanish and English elite teams. *Journal of Sports Sciences,* 34(24), 2195-2204.

Fernandez-Navarro, J., Fradua, L., Zubillaga, A., McRobert, A.P. (2018). Influence of contextual variables on styles of play in soccer. *International Journal of Performance Analysis in Sport*, 18(3), 423-436.

Feuerhake, U. (2016). Recognition of repetitive movement patterns - the case of football analysis. *International Journal of Geo-Information*, 5(11), 1-20.

Feuerhake, U., Sester, M. (2013). Mining group movement patterns. In Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographical Information Systems (pp.520-523). Florida, FL, U.S.A.: ACM.

Figueroa, P., Leite, N.J., Barros, R.M.L. (2006). Tracking soccer players aiming their kinematical motion analysis. *Computer Vision and Image Understanding*, 101(2), 122-135.

Fréchet, M. (1960). Sur quelques points du calcul fonctionnel. *Rendiconti del Circolo Matematico di Palermo*, 22, 1-74.

Gomez, G., Herrera Lopez, P., Link, D., Eskofier, B. (2014). Tracking of ball and players in beach volleyball videos. *PLOS One*, 9(11), 1-19.

Gonçalves, B.V., Figueira, B.E., Maçãs, V., Sampaio, J. (2014). Effect of player position on movement behaviour, physical and physiological performances during an 11-a-side football game. *Journal of Sports Sciences*, 32(2), 191-199.

Grunz, A., Memmert, D., Perl, J. (2012). Tactical pattern recognition in soccer games by means of special self-organizing maps. *Human Movement Science*, 31(2), 334-343.

Gudmundsson, J., Wolle, T. (2014). Football analysis using spatio-temporal tools. *Computers, Environment and Urban Systems*, 47, 16-27.

Haugen, T., Tønnessen, E., Hisdal, J.,, Seiler, S. (2013). The role and development of sprinting in soccer. *International Journal of Sports Physiology and Performance*, 9(3), 432-441.

Hewitt, A., Greenham, G., Norton, K. (2016). Game style in soccer: what is it and can we quantify it? *International Journal of Performance Analysis in Sport*, 16(1), 355-372.

Jain, A.K., Duin, R.P.W., Mao, J. (2000). Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 22(1), 4-37.

James, N. (2006). The role of notational analysis in soccer coaching. *International Journal of Sports Science & Coaching*, 1(2), 185-198.

Jeffreys, I. (2008). Movement training for field sports: soccer. *Strength and Conditioning Journal*, 30(4), 19-27.

Jonsson, G.K., Anguera, M.T., Sánchez-Algarra, P., Olivera, C., Campanico, J., Castañer, M., Torrents, C., Dinusová, M., Chaverri, J., Camerino, O., Magnusson, M.S. (2010). Application of t-pattern detection and analysis in sports research. *The Open Sports Sciences Journal*, 3, 95-104.

Kang, C.H., Hwang, J.R., Li, K.J. (2016). Trajectory analysis for soccer players. In Proceedings of the 6th IEEE International Conference on Data Mining (pp. 377-381). Hong Kong, China.

Kelley, J., Chopin, S., Goodwill, S., Haake, S. (2010). Validation of a live, automatic ball velocity and spin rate finder in tennis. *Procedia Engineering*, 2(2), 2967-2972.

Kelly, D.M., Drust, B. (2009). The effects of pitch dimensions on heart rate responses and the technical demands of small-sided soccer games in elite players. *Journal of Science and Medicine in Sport*, 12(1), 475-479.

Lago-Ballesteros, J., Lago-Peñas, C., Rey, E. (2012). The effect of playing tactics and situational variables on achieving score-box possessions in a professional soccer team. *Journal of Sports Sciences*, 30, 1455-1461.

Lago-Peñas, C., Gómez-Ruano, M., Yang, G. (2017). Styles of play in professional soccer: an approach of the Chinese Soccer Super League. *International Journal of Performance Analysis in Sport*, 17(6), 1073-1084.

Lames, M., Ertmer, J., Walter, F. (2010). Oscillations in football—order and disorder in spatial interactions between the two teams. *International Journal of Sport Psychology*, 41(4), 85.

Laube, P., Van Kreveld, M., Imfeld, S. (2004). Finding REMO – Detecting relative motion patterns in geospatial lifelines. In P.D. Fisher (Ed.), Developments in spatial data handling, Proceedings of the 11th international symposium on spatial data handling (pp. 201-214). Berlin, Heidelberg, Germany: Springer.

Lee, J.G., Han, J., Whang, K.Y. (2007). Trajectory clustering: a partition-and-group framework. In Proceedings of the ACM IGMOD international conference on management of data (pp. 593-604). Beijing, China.

Leite, N.M., Leser, R., Gonçalves, B., Calleja-Gonzalez, J., Bacca, A., Sampaio, J. (2014). Effect of defensive pressure on movement behaviour during an under-18 basketball game. *International Journal of Sports Medicine*, 35(9), 743-748.

Leser, R., Hoch, T., Tan, X., Moser, B., Kellermayr, G., Baca, A. (2018). Finding efficient strategies in 3-versus-2 small-sided games of youth soccer players. *Kinesiology: International Journal of Fundamental and Applied Kinesiology*, 51(1), 1-2.

Link, D. (2018). Sports Analytics. How (commercial) sports data create new opportunities for sports science. *German Journal of Exercise and Sports Research*, 48(1), 13-25.

Magnusson, M.S. (1996). Hidden real-time patterns in intra- and inter-individual behaviour. *European Journal of Psychological Assessment*, 12(2), 112-123.

Magnusson, M.S. (2000). Discovering hidden time patterns in behaviour: T-patterns and their detection. *Behavior Research Methods, Instruments, & Computers*, 32(1), 93-110.

Meilgaard, M., Civille, G.V., Carr, B.T. (2007). Sensory evaluation techniques. Boca Raton, FL, U.S.A.: CRC Press.

Memmert, D., Lemmink, K.A.P.M., Sampaio, J. (2017). Current approaches to tactical performance analysis in soccer. *Sports Medicine*, 47(1), 1-10.

Memmert, D., Raabe, D. (2018). Data Analytics in Football: Positional Data Collection, Modelling and Analysis. London, U.K.: Routledge.

Nanni, M., Pedreschi, D. (2006). Time-focused clustering of trajectories of moving objects. *Journal of Intelligent Information Systems*, 27, 267-289.

Nicholls, S.B., Worsfold, P.R. (2016). The observational analysis of elite coaches within youth soccer: the importance of performance analysis. *International Journal of Sports Science & Coaching*, 11(6), 825-831.

Niu, Z., Gao, X., Tian, Q. (2012). Tactical analysis based on real-world ball trajectory in soccer video. *Pattern Recognition*, 45, 1937-1947.

Perl, J., Grunz, A., Memmert, D. (2013). Tactics Analysis in Soccer–An Advanced Approach. *International Journal of Computer Science in Sport*, 12, 33-44.

Peryam, D.R., Swartz, V.W. (1950). Measurement of sensory differences. *Food Technology*, 4(1), 390-395.

Pérez-Turpin, J.A. Cortell-Tormo, J.M., Suárez-Llorca, C., Chinchilla-Mira, J.J., Cejuela-Anta, R. (2009). Gross movement patterns in elite female beach volleyball. *Kinesiology*, 41(2), 212-219.

Rafael, M.D., Antonio, N., Asier, Z., Luis, F. (2017). Influence of tactical variables on attacking actions in elite soccer. *Kinesiologia Slovenica*, 23(3), 32-38.

Randers, B.R., Orntoft, C., Hagman, M., Nielsen, J.J., Krustrup, P. (2017). Movement pattern and physiological response in recreational small-sided football - effect of number of players with a fixed pitch size. *Journal of Sports Sciences*, 36(13), 1-8.

Rein, R., Memmert, D. (2016). Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. *SpringerPlus*, 5(1), 1410.

Sampaio, J., Maçãs, V. (2012). Measuring tactical behaviour in football. *International Journal of Sports Medicine*, 33(5), 395-401.

Sarmento, H., Anguera, T., Campaniço, J., Leitão, J. (2010). Development and validation of a notational system to study the offensive process in football. *Medicina*, 46(6), 401.

Sarmento, H., Marcelino, R., Anguera, M.T., Campanico, J., Matos, N., Leitao, J.C. (2014). Match analysis in football: a systematic review. *Journal of Sports Sciences*, 32(20), 1831-1843.

Sasha, D., Al-Masoudi, F., Stein, M., Schreck, T., Keim, D.A., Adrienko, G., Janetzko, H. (2017). Dynamic visual abstraction of soccer movements. *Computer Graphics Forum*, 36(6), 305-315.

Seweryniak, T., Mroczek, D., Lukasik, L. (2013). Analysis and evaluation of defensive team strategies in women's beach volleyball - an efficiency-based approach. *Human Movement*, 14(1), 48-55.

Sha, L., Lucey, P., Yue, Y., Carr, P., Rohlf, C., Matthews, I. (2016). Chalkboarding: a new spatiotemporal query paradigm for sports play retrieval. In Proceedings of the 21st International Conference on Intelligent User Interfaces (pp. 356-364). Sonoma, CA, U.S.A.: ACM

Shao, L., Sacha, D., Nelder, B., Stein, M., Schreck, T. (2016). Visual-interactive search for soccer trajectories to identify interesting game situations. In Proceedings of the Conference on Visualisation and Data Analysis (pp. 1-10). San Jose, CA, U.S.A.

Sweeting, A.J., Aughey, R.J., Cormack, S.J., Morgan, S. (2017). Discovering frequently recurring movement sequences in team-sport athlete spatiotemporal data. *Journal of Sports Sciences*, 35(24), 2439-2445.

Stone, H., Bleiblaum, R.N., Thomas, H.A. (2012). Sensory evaluation practices. London, U.K.: Elsevier.

Tenga, A., Holme, I., Ronglan, L.T., Bahr, R. (2010). Effect of playing tactics on achieving score-box possessions in a random series of team possessions from Norwegian professional soccer matches. *Journal of Sports Sciences*, 28(3), 245-255.

Tenga, A., Larsen, O. (2003). Testing the validity of match analysis to describe playing styles in football. *International Journal of Performance Analysis in Sport*, 3(2), 90-102.

Vilar, L., Esteves, P.T., Travassos, B., Passos, P., Lago-Penãs, C., Davids, K. (2014).Varying numbers of players in small-sided soccer games modifies action opportunities during training. *International Journal of Sports Science & Coaching*, 9(5), 1007-1018.

Winter, C., Pfeiffer, M. (2016). Tactical metrics that discriminate winning, drawing and losing teams in EUFA Euro 2012®. *Journal of Sports Sciences*, 34(6), 486-492.

Zhang, P.D., Beernaerts, J., Zhang, L., Van de Weghe, N. (2016). Visual exploration of match performance based on football movement data using the continuous triangular model. *Applied Geography*, 76, 1-13.

"Nothing is more revealing
than movement"

– Martha Graham –

7.

USE CASE 2: DYNAMIC GAIT ANALYTICS

Parts of this chapter were modified from: BEERNAERTS, J., DERIE, R., NGUYEN, B., VANSTEENKISTE, P., DE BAETS, B., DECONINCK, F.J.A., LENOIR, M., DE CLERCQ, D., VAN DE WEGHE, N. (2018). Assessing the potential of the qualitative trajectory calculus to detect gait pathologies: a case study of children with development coordination disorder. *Computer Methods in Biomechanics and Biomedical Engineering*, 1-7.

In this Chapter, the use of QTC for a more detailed movement analyses of the human body is investigated, aiming at providing an answer to research questions **R3** & **Rd.** For this, we investigate whether the proposed method for dynamic sports analytics can be used to perform gait analysis. Gait analysis can be defined as the study of human motion, which is conducted by measuring and analysing the human body movements, body mechanics and muscle activity (Cappozzo, 1984). While gait analysis could be conducted for human body movements in any sport, difficulties regarding data collection limit current uses to sports that can be performed in controlled environments (*i.e.* indoor laboratories). Because of this, gait analysis is mostly used for analysing walking and running movements. Nevertheless, analysing walking and running is of big importance for sports analytics, as walking and running are prevalent in most of the popular sports. In this use case, we demonstrate that the proposed method for dynamic sports analytics can be used for detecting subtle differences in walking behaviour.

The first section of this chapter gives an introduction to the use case. Given the novelty of the proposed method for dynamic sports analytics in the domain of gait analytics, the overview of the relevant literature is quite concise. For this reason, it is included in Section 7.1, rather than presented separately. The second section includes a description of the used dataset and set-up of the use case. Section 7.3 introduces the approach that is used for validating the results of the gait analysis for each combination of markers. The results are presented in the Section 7.4, together with the validation. To conclude, Section 7.5 includes a discussion of the results and their value for the domain of gait analysis. An evaluation of the posed research questions is not included in this chapter, as this is presented comprehensively for all research questions and use cases in Chapter 9.

7.1 Introduction

Nixon *et al.* (2006) demonstrated that walking patterns are unique for every individual person. Gait analysis can therefore be used both for studying the specific walking behaviour of individuals, as well as for the detection of mutual movement characteristics of groups of people. Gait analysis has traditionally been studied using a wide range of approaches including kinetic and kinematic analyses, and the measurement of muscle activity (Hsiao-Wecksler *et al.*, 2010; König *et al.*, 2016; McGinley *et al.*, 2009). However, for clinical purposes, complex datasets are often difficult to interpret because of the large number of quantitative indicators. Therefore, there have been many attempts to summarize these measures into a concise index that could serve as a single measure for the quality of the gait pattern. Examples are the normalcy index (PCA on 3 spatio-temporal and 13 kinematic gait analysis parameters), the hip flexor index (PCA on 5 kinematic and 5 kinetic gait analysis parameters), the gait deviation index (kinematic comparison of 15 features of pelvis and hip (3D), knee and ankle (2D) and foot progression with a normal gait pattern) and the gait profile score (Cimolin & Galli, 2014). Despite the fact that these summary measures have been applied successfully for the detection of gait pathologies, gait analyses in a clinical setting are still often performed by subjective qualitative human observers (Chen *et al.*, 2016).

The main advantage of applying the proposed method for dynamic sports analytics to gait analysis is that it does not require the interpretation of such complex datasets (e.g. a combination of 3D kinematic, kinetic, and often EMG time series) and/or these numerous quantitative indicators. Rather than giving a series of quantitative indicators, the method can give a single value (*i.e.* distance or similarity) indicating whether two gait patterns movements are similar or not. This could aid human observers into detecting gait pathologies.

In the current use case, we investigate whether QTC is able to detect differences in the gait pattern of children with and without the Developmental Coordination Disorder (DCD). DCD is a neurodevelopmental disorder characterized by reduced ability to employ fundamental motor skills (American Psychiatric Association, 2013). Using quantitative methods (see first paragraph of this section), subtle differences in gait between children with and without DCD have been found (Deconinck *et al.*, 2006; Deconinck *et al.*, 2010), especially when the task demands are high (e.g. when walking in the dark or over uneven terrain). However, distinguishing children with DCD from Typically Developing children (TD) based on their gait pattern has been proven to be very difficult (Cherng *et al.*, 2009; Du *et al.*, 2015). Besides the availability of a dataset from previous research (Deconinck *et al.*, 2006), this difficulty was the main motivation for selecting this topic as use case in this dissertation.

7.2 Dataset and analysis set-up

In this section, the dataset and the set-up of the experiment are discussed, giving a detailed overview of the conducted research.

7.2.1 Dataset

For this use case, the dataset from a quantitative study of differences in walking behaviour between children with and without DCD (Deconinck *et al.*, 2006) is used. This spatio-temporal dataset contains (x,y,z,t)-coordinates of ten children with DCD and ten gender-matched TD children walking on a motor driven treadmill. For each child, walking speed was scaled to the leg length according to the Froude number (Fr = $v^2/g.L$ with v = walking velocity, g = gravitational acceleration, L = leg length).

When walking at an equal Froude number, one can assume that walking occurs with dynamic similarity (*i.e.* with similar length, timing, frequency, velocity and forces; Zatiorsky *et al.*, 1994). The Froude number was set to 0.15, resulting in a mean walking velocity of 0.85 m/s for both the DCD and the TD population. All the children of the DCD group met the DSM IV criteria for DCD (American Psychiatric Association, 2013). For each of the children, 14 markers (Figure 7-1) were tracked at a frequency of 240Hz. For a more extensive description of the dataset, we refer to the study of Deconinck et al. (2006).

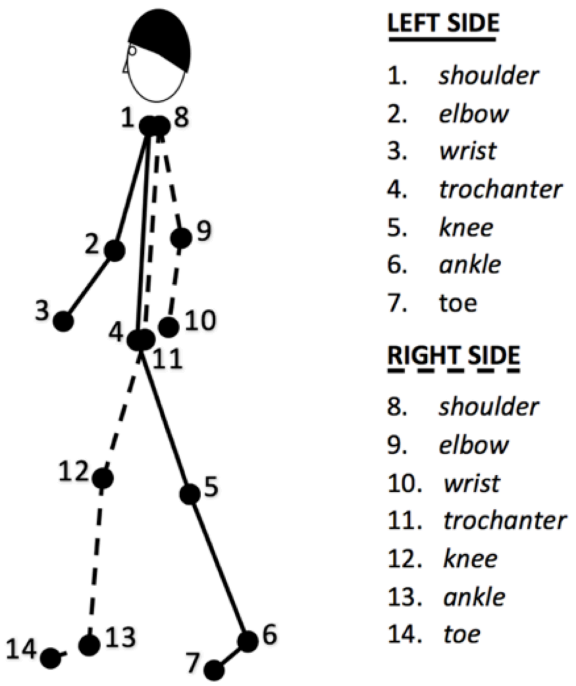


Fig. 7-1. Lateral representation of the markers in the dataset. Note that only the right side of the body was analysed (dashed line).

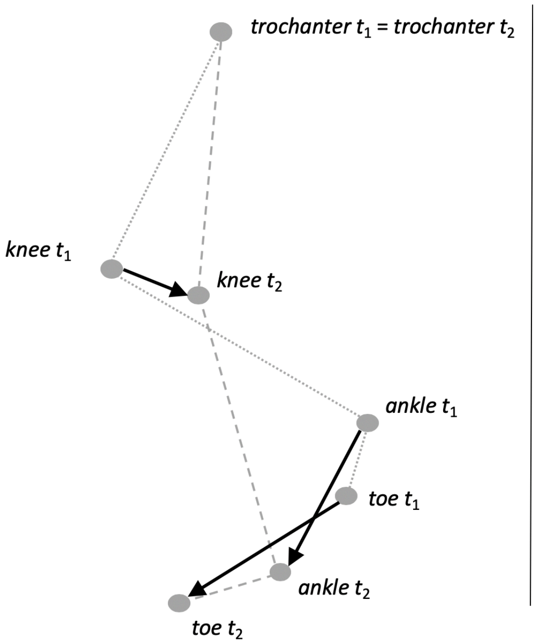
Before performing the analysis, several selections and adaptations are made to the original dataset. First, due to incomplete marker data, one subject of each population is excluded, resulting in a dataset of 18 children (9 DCD and 9 TD). Second, assuming symmetry between both of the lateral sides of the body (Whitall & Caldwell, 1992), only the right side of the body is taken into consideration. This reduces the original set of markers to 7 unilateral markers (markers 8 till 14 in Figure 7-1). Third, to ensure all stance phases, hereinafter referred to as ‘steps’, are complete and uninterrupted, the first and the last steps are excluded. This results in four consecutive steps of the right foot for each participant for the analysis. Foot strikes and toe-offs are thereby determined as described by Donker *et al.* (2001) and defined as respectively the timing of maximal forward position of the ankle markers, and maximal backward position of the toe markers.

In summary, for each of the 18 children (9 DCD and 9 TD), four consecutive steps are extracted resulting in a dataset of 72 steps. With 7 markers (Figure 7-1) available, all different subsets of 2, 3, 4, 5, 6, or 7 markers can be used for the qualitative comparison of the different steps in order to discern between the DCD and TD groups.

7.2.2 Analysis set-up

When choosing a set of markers on the body as MPOs, QTC can be used to make a qualitative description of a person's walking pattern by describing the relative movements of this set of markers, conform the methodology for dynamic sports analytics described in Chapter 3. The most basic QTC-variant (QTC_{B21}) is chosen for this analysis because of the absence of previous research of QTC analysis in this domain and the relatively large number of markers (7). Furthermore, it can be assumed that for an expert it is easier to compare QTC-relations of a less detailed variant with the results of classic gait analyses (see Section 7.1). For each time interval, a QTC_{B21}-relation can be constructed for each pair of markers, which are gathered in a QTC_{B21}-matrix. This is exemplified in Figure 7-2 for a set of markers containing the *trochanter*, *knee*, *ankle* and *toe*. When describing movements over multiple time intervals, the QTC_{B21}-matrices can be put into a sequence. As such, for each step of each person, a QTC_{B21}-matrix sequence representation can be constructed for a given set of markers (see Section 3.1). Since the steps have different temporal lengths, the corresponding QTC-matrix sequence representations will differ in length (number of QTC_{B21}-matrices) accordingly. Next, the distance between each pair of steps (for a given combination of markers) is calculated, by comparing their QTC_{B21}-matrix sequences, which was described rigorously in Section 3.7. This Levenshtein distance metric is used to cope with the differences in lengths of the different steps as well as with differences in pace during (parts of) the steps.

No static points are taken into account, as all the data was gathered on a treadmill that had a fixed location and orientation with respect to the cameras that were used for data logging. Furthermore, no angular tolerance for the 0-character is used (α =0°), given the limitations mentioned in Section 3.5. When comparing two steps, it is important to compare movement of corresponding body parts (e.g. the movement of the *ankle*-marker from step 1 should be compared with the movement of the *ankle*-marker of step 2). For this reason, no (marker) permutations are used for the distance calculation between the different steps.



LEGEND

Line connecting markers at t_1

Line connecting markers at t_2

Vector showing marker movement between t_1 and t_2

QTC-MATRIX

$T(t_1 - t_2)$	*trochanter*	*knee*	*ankle*	*toe*
trochanter		**0 +**	**0 +**	**0 +**
knee			**- +**	**- +**
ankle				**- +**
toe				

Fig. 7-2. Positions of the markers (*trochanter, knee, ankle and toe*) on the leg at two timestamps t_1 and t_2 (left) and the corresponding QTC_{B21}-matrix.

7.3 Selecting the marker combination for comparing of DCD and TD walking patterns

7.3.1 Classifier

In the dataset, 7 markers are available, although it is unknown which of these might be useful for this specific gait analysis. We therefore aim at finding the optimal combination of markers that best allows us to correctly classify a(n) (unknown) step of a child as DCD or TD, based on the distances between the QTC_{B21}-representations of all steps for that marker combination. For this task, the *k*-nearest-neighbour (*k*-NN) classifier is chosen as a baseline classifier due to its simplicity and effectiveness for classification problems (Cover & Hart, 1967). First, *k*-NN makes no assumptions about the training data, making it suitable for many different problem areas. Second, *k*-NN is a distance-based classifier, which means that only distances are required in order to perform the classification. The idea of *k*-NN is quite simple and amounts to assigning to an unlabelled object the majority label among its *k* nearest neighbours in the training set. To choose the best model, only the hyper-parameter *k* needs to be tuned. It is well known that *k*-NN is sensitive to the number of neighbours *k* (Cover & Hart, 1967). Since in this research there are only two classes (TD and DCD), odd values of *k* (*i.e.* 1, 3, …, 13) are used to avoid tied votes.

7.3.2 Performance measurement

Let $\{(x_p, y_p)\}_{p=1}^{n}$ denote a set of children and their corresponding class label (DCD or TD), where $n = 18$ is the number of children. Children belonging to the TD group are labelled as '1' and children belonging to the DCD group are labelled as '0'. Since the training set contains a rather small number of children, leave-one-out cross-validation was used to assess the performance of the classifier. To be more specific, one child at a time is used as a test example and the other children are used as training examples. In this experiment, two performance metrics are used in order to evaluate the methodology: the error at step level and the error at individual level.

Error at step level (E_{step}): This error is defined as the error of misclassifying a child using information from just a single step of that child. For each child, distances between only one step (ignoring the other three steps of that child) and all steps of the other children are used. Since each child has 4 steps, this results in 4 predictions (TD/DCD) with corresponding probabilities. Let $f(x_p,i)$ denote the prediction probability of child x_p being a normal child (TD) using step i, which is computed as the proportion of nearest neighbours used for k-NN that are labelled TD. A child is classified as TD if the prediction probability of TD is larger than or equal to that of being DCD, *i.e.* $f(x_p,i) \geq 0.5$. Therefore, the error at step level is formalized as follows:

$$E_{step} = \frac{\sum_{p=1}^{n} \sum_{i=1}^{m} \mathbf{1}\left(\mathbf{1}\left(f\left(x_p,i\right) \geq 0.5\right) \neq y_p\right)}{nm}$$

where the equation above is the indicator function which results in a value of '1' if its argument is true, and '0' otherwise, and $m = 4$ denotes the number of steps measured for each child. A misclassification occurs when the predicted group of a child is TD (*i.e.* $f(x_p,i) \geq 0.5$) and (s)he has DCD (*i.e.* $y_p = 0$) or on the contrary when the child is predicted as DCD (*i.e.* $f(x_p,i) < 0.5$) and (s)he is part of the control group (TD) (*i.e.* $y_p = 1$).

Error at individual level ($E_{individual}$): This error is defined as the error of misclassifying a child using all steps of that child. For classifying a child as TD or DCD at the individual level, it combines the different predictions at the step level. Based on these predictions, a child is classified as normal if the majority of his/her predictions at step level predicts him/her as normal, *i.e.* if $\frac{\sum_{i=1}^{m} f(x_p,i)}{4} \geq 0.5$. Formally, the error at individual level is computed as:

$$E_{individual} = \frac{\sum_{p=1}^{n} \mathbf{1}\left(\mathbf{1}\left(\frac{\sum_{i=1}^{m} f(x_p,i)}{4} \geq 0.5\right) \neq y_p\right)}{n}$$

7.3.3 Selecting the best marker combination

For each marker combination, both E_{step} and $E_{individual}$ are calculated. A good combination of markers should have a low classification error, meaning that k-NN is capable of discriminating between DCD and TD children.

7.4 Results

Different marker combinations are available to analyse the gait of children with QTC. The aim is to find the best marker combination to discern between the DCD and TD groups. The best results of the experiment are shown in Table 7-1. This table includes the best leave-one-out error according to $E_{individual}$ using different numbers of markers and different numbers of neighbours (k). E_{step} is typically larger than $E_{individual}$, which can be explained by the fact that $E_{individual}$ is computed using all steps from each child, making it more reliable than E_{step}. The best $E_{individual}$ (equal to 0.167) is obtained using a combination of 3, 4 or 5 markers, shown in boldface in the table. An overview of the results for all marker combinations is included in Appendix 4 at the end of this dissertation.

The best marker combination consists of the three markers *ankle*, *toe*, and *trochanter*. When adding the marker *shoulder*, or the markers *shoulder* and *elbow* to this combination, the resulting marker combinations of 4 and 5 markers, respectively, keeps the same $E_{individual}$. In the combination of 5 markers, however, we report a slight increase in E_{step}. Based on the experimental results, we can conclude that there are two best marker combinations: the three-marker combination *ankle*, *toe*, and *trochanter* and the four-marker combination *ankle*, *toe*, *trochanter*, and *shoulder*. Both combinations have identical $E_{individual}$ and E_{step} which were reported as the lowest for all the marker combinations. Using these marker combinations as MPOs, the dynamic QTC analysis was able to classify 83.3% of the subjects correctly to either the DCD group or the TD group. Due to its simplicity, the three-marker combination is preferred over the four-marker combination, as it means that similar results are achieved with a simpler model and thus less calculations.

Table 7-1. Overview of the best leave-one-out errors using different marker combinations. The best results are highlighted in boldface, with ✓ denoting the used markers. Note that all marker combinations are tested but only the best ones are shown in this table. For each combination of markers in the table, the number of markers (#), $E_{individual}$, E_{step} and the number of used neighbours (k) in k-NN are shown.

#	$E_{individual}$	E_{step}	k	Markers selected						
				ankle	*elbow*	*knee*	*shoulder*	*toe*	*trochanter*	*wrist*
2	0.222	0.347	5				✓			✓
		0.361	7					✓		✓
		0.444	9	✓				✓		
		0.361	13	✓				✓		
3	**0.167**	**0.333**	3	✓				✓	✓	
4	**0.167**	**0.333**	3	✓			✓	✓	✓	
5	**0.167**	0.389	3	✓	✓		✓	✓	✓	
6	0.278	0.431	3	✓	✓	✓		✓	✓	✓
		0.417	5	✓	✓	✓	✓		✓	✓
		0.347	5	✓		✓	✓	✓	✓	✓
		0.389	7	✓	✓	✓		✓	✓	✓
		0.389	7	✓	✓	✓	✓		✓	✓
		0.333	7	✓		✓	✓	✓	✓	✓
		0.361	11	✓		✓	✓	✓	✓	✓
		0.417	13	✓	✓	✓	✓	✓	✓	✓
7	0.278	0.403	7	✓	✓	✓	✓	✓	✓	✓
		0.389	9	✓	✓	✓	✓	✓	✓	✓

Table 7-2 displays the subject specific classifications based on this three-marker combination. It shows, for each four steps separately as well as for the combination of the four steps, whether the child is correctly classified as TD or DCD. From this table it can be seen that that children 5 and 6 are incorrectly classified (DCD) for each of their individual steps and therefore also for the combination of all their steps. For children 1, 4 and 8, however, the overall classification is correct (TD) as only one or two of their steps resulted in an incorrect classification (DCD). Children 2, 3, 7 and 9 are correctly classified as TD using each of their steps. For the children with DCD, only child 12 is classified correctly for all separate steps. Children 10, 11, 13, 14, 16, 17 and 18 all have one or more incorrectly classified step, but are classified correctly (DCD) based on all of their four steps. Child 15 on the other hand, is the only child with DCD that is classified incorrectly (TD).

Table 7-2. Subject specific classifications based on the best model (*ankle*, *toe* and *trochanter*, see Table 1) for the different steps separately and all four steps combined. Correct classifications are annotated with '✓' ,incorrect classifications are annotated with '–'.

Child number	Real class child	Child correctly classified using				
		Step 1	*Step 2*	*Step 3*	*Step 4*	All four steps
1	TD	✓	-	✓	✓	✓
2	TD	✓	✓	✓	✓	✓
3	TD	✓	✓	✓	✓	✓
4	TD	✓	✓	✓	-	✓
5	TD	-	-	-	-	-
6	TD	✓	-	-	-	-
7	TD	✓	✓	✓	✓	✓
8	TD	-	-	✓	✓	✓
9	TD	✓	✓	✓	✓	✓
10	DCD	✓	✓	✓	-	✓
11	DCD	✓	✓	✓	-	✓
12	DCD	✓	✓	✓	✓	✓
13	DCD	✓	✓	-	✓	✓
14	DCD	✓	-	✓	-	✓
15	DCD	-	-	-	✓	-
16	DCD	✓	-	-	✓	✓
17	DCD	✓	✓	-	✓	✓
18	DCD	✓	-	-	✓	✓

7.5 Discussion

In this section, the importance of the results for the domain of gait analysis are discussed. An evaluation of the use case for the sports analytics domain and its value for answering the respective research questions can be found in Chapter 9.

The results of Section 7.4 show that a classification accuracy of 83.3% can be achieved for distinguishing between children with and without DCD, using only three markers as MPOs (*ankle*, *toe*, *trochanter*). For this (gait) analysis, it is important to define the optimal marker combination, as each added MPO to this combination adds more noise to the QTC-description of the steps. Consequently, it is important to identify the optimal marker combination for each application of the QTC dynamic sports analytics method. With these markers, only the 'essential' information is analysed, while irrelevant information is ignored.

The results also show that including more steps increases the predictive ability of the analysis ($E_{individual} < E_{step}$). This is probably due to the fact that children with DCD typically have difficulties to produce consistent movement patterns (Deconinck *et al.*, 2006; Rosengren *et al.*, 2009). Unfortunately, based on the current dataset it is not possible to test whether more steps would result in an even better discrimination, and whether there is an optimal number of steps to be included for a QTC-analysis of DCD gait pathology. Although there are subtle kinetic and kinematic differences between overground and treadmill walking (Riley, 2008), Tesio *et al.* (2017) concluded that gait analysis on a treadmill seems to be a promising alternative to conventional gait analysis of children. More research, however, is needed to generalise the results of our study to overground walking.

That the proposed method gives the best results using a relatively low number of markers makes the results easier to interpret and requires less computational time and resources. Whereas traditional quantitative approaches for gait analysis usually provide more detailed results as more measures are included, the proposed method does not seem to benefit from using more markers. Since the volume of data produced by quantitative three-dimensional gait analyses has been described as a potential obstacle for its clinical use (Cimolin & Galli, 2014), the proposed method might be a more feasible alternative. Although gait pathology for children with DCD has been considered as very challenging (Wilmut *et al.*, 2016), the dynamic sports analytics method was already able to classify 83.3% of the subjects to the correct group. In comparison, data of gait variability measured by trunk accelerometers, using the results from a ROC analysis from the short-term local dynamic stability as a classifier, showed a classification accuracy ranging between 78% and 89% in detecting the gait pattern from DCD children compared to typically developing children (Speedtsberg *et al.*, 2018). Unfortunately, other studies on DCD usually only describe differences between a DCD and a non-DCD population, without testing the predictive value of their findings. Although the classification accuracy of 83.3% seems high compared to literature (Woodruff et al. 2002, Speedtsberg et al. 2018), it does not suffice for a clinical test. By using more than four steps per individual, the overall classification accuracy could be enhanced, yet this remains an assumption. Nevertheless, the advantage of the proposed method for dynamic sports in this domain is that it could be easily used to analyse walking or running patterns for different gait pathologies.

REFERENCES

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5^{th} ed.), Washington, DC, U.S.A.

Cappozzo, A. (1984). Gait analysis methodology. *Human Movement Science*, 3, 27-50.

Chen, S., Lach, J., Member, S., Lo, B., Member, S. (2016). Towards pervasive gait analysis with wearable sensors: A systematic review. *IEEE Journal of Biomedical and Health Informatics*, 20(6), 1521-1537.

Cherng, R.J., Liang, L.Y., Chen, Y.J., Chen, J.Y. (2009). The effects of a motor and a cognitive concurrent task on walking in children with developmental coordination disorder. *Gait and Posture*, 29(2), 204-207.

Cimolin, V., Galli, M. (2014). Summary measures for clinical gait analysis: A literature review. *Gait and Posture*, 39(4), 1005-1010.

Cover, T.M., Hart, P.E. (1967). Nearest neighbour pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21-27.

Deconinck F.J.A., De Clercq, D., Savelsbergh, G.J.P., Lenoir, M. (2010). Balance problems during obstacle crossing in children with Developmental Coordination Disorder. *Gait and Posture*, 32(3), 327-331.

Deconinck, F.J.A., De Clercq, D., Savelsbergh, G.J.P., Van Coster, R., Oostra, A., Dewitte, G., Lenoir, M. (2006). Differences in gait between children with and without developmental coordination disorder. *Motor Control*, 10(2), 125-42.

Donker, S.F., Beek, P.J., Wagenaar, R.C., Mulder, T. (2001). Coordination between arm and leg movements during locomotion. *Journal of Motor Behavior*, 33(1), 86-102.

Du, W., Wilmut, K., Barnett, A.L. (2015). Level walking in adults with and without Developmental Coordination Disorder: An analysis of movement variability. *Human Movement Science*, 43(1), 9-14.

Hsiao-Wecksler, E.T., Polk, J.D., Rosengren, K.S., Sosnoff, J.J., Hong, S. (2010). A review of new analytic techniques for quantifying symmetry in locomotion. *Symmetry*, 2(2), 1135-1155.

König, N., Taylor, W.R., Baumann, C.R., Wenderoth, N., Singh, N.B. (2016). Revealing the quality of movement: A meta-analysis review to quantify the thresholds to pathological variability during standing and walking. *Neuroscience and Biobehavioral Reviews*, 68(1), 111-119.

McGinley, J.L., Baker, R., Wolfe. R., Morris, M.E. (2009). The reliability of three-dimensional kinematic gait measurements: A systematic review. *Gait and Posture*, 29(3), 360-369.

Nixon, M.S., Tan, T., Chellappa, R. (2006). Human identification based on gait. New York, NY, U.S.A.: Springer.

Riley, P.O., Dicharry, J., Franz, J., Della, C.U., Wilder, R.P., Kerrigan, D.C. (2008). A kinematics and kinetic comparison of overground and treadmill running. *Medicine & Science in Sports & Exercise*, 40(6), 1093-1100.

Rosengren, K.S., Deconinck, F.J.A., Diberardino, L.A., Polk, J.D., Spencer-Smith, J., De Clercq, D., Lenoir, M. (2009). Differences in gait complexity and variability between children with and without Developmental Coordination Disorder. *Gait and Posture*, 29(2), 225-229.

Speedtsberg, M.B., Christensen, S.B., Stenum, J., Kallemose, T., Bencke, J., Curtis, D.J., Jensen, B.R. (2018). Local dynamic stability during treadmill walking can detect children with development coordination disorder. *Gait Posture*, 59(1), 99-103.

Tesio, L., Malloggi, C., Portinaro, N.M., Catino, L., Lovecchio, N., Rota, V. (2017). Gait analysis on force treadmill in children: comparison with results from ground-based force platforms. *International Journal of Rehabilitation Research*, 40(4), 315-324.

Whitall, J., Caldwell, G.E. (1992). Coordination of symmetrical and asymmetrical human gait. *Journal of Motor Behavior*, 24(4), 339-353.

Wilmut, K., Du, W., Barnett, A.L. (2016). Gait patterns in children with Developmental Coordination Disorder. *Experimental Brain Research*, 234(6), 1747-1755.

Woodruff, S.J., Bothwell-Meyers, C., Tingley, M., Albert, W.J. (2002). Gait pattern classification of children with developmental coordination disorder. *Adapted Physical Activity Quarterly*, 19(3), 378-391.

Zatsiorsky, V.M., Werner, S.L., Kaimin, M.A. (1994). Basic kinematics of walking: Step length and step frequency: A review. *Journal of Sports Medicine and Physical Fitness*, 34, 109-134.

"The first 90 minutes are the most important"

– Bobby Robson –

8.

USE CASE 3: STATIC FOOTBALL ANALYTICS

Parts of this chapter were modified from: BEERNAERTS, J., DE BAETS, B., LENOIR, M., DE MEY, K., VAN DE WEGHE, N. (2018). Analysing team formations in football with the static qualitative trajectory calculus. In P. Pezarat-Correia, J.P. Vilas-Boas, O. Rivera & J. Cabri (Eds), Proceedings of the 6th International Congress on Sport Sciences Research and Technology Support (pp.15-22). Seville, Spain: Scitepress.

In this Chapter, the use of the proposed method for static sports analytics for the analysis of team tactics is investigated. The aim of this chapter is to provide an answer to research question **R2** and more specifically research questions **Ra**, **Rb** and **Rc**. For this a use case is designed, evaluating whether the proposed method for static sports analytics can be used to analyse team formation in football. The use case consists of a series of football examples and one football experiment, based on real professional football matches. The first section presents a literature overview of team tactic and team formation analysis in sports and football more specifically. In Section 8.2, the used datasets and analysis set-up of both the examples as well as the experiment are discussed. Section 8.3 includes a series of team formation analysis examples in football. The more detailed experiment is presented in the fourth section, including a validation of the experiments results based on popular media match reports. Section 8.5 includes a discussion presenting advantages and limitations of the proposed method for team formation analysis in football, based on the results of the examples and the experiment. An evaluation of the posed research questions is not included in this chapter, as this is presented comprehensively for all research questions and use cases in Chapter 9.

8.1 Team formation analysis in sports

This section presents a literature overview of team formation analysis in sports. Since almost all team formation analysis methods, regardless of the sport, use positional data of the players, we start by giving a brief but focused overview of the state-of-the-art of team formation analysis in popular team ball sports, before providing a broader overview of the domain for football. For a more general overview of all different sports analytics methods in football, we refer to the works of Rein & Memmert (2016) and Sarmento *et al.* (2014) and Memmert *et al.* (2017).

8.1.1 Team ball sports

In American football, Atmosukarto *et al.* (2013) did efforts for the automatic recognition of offensive team formations, which they defined as “The spatial configuration of a team’s players before a play starts”. Their method automatically detects when one of five reference offensive team formations is achieved during the game. The big difference with football, however, is that a football game is more fluent and dynamic, thus team formations tend to change more during the course of the game. Team formation analysis in volleyball has been conducted by Jäger & Shöllhorn (2012). Because of the distinct separation of a volleyball game in separate rallies, Jäger & Shöllhorn used the positions of the players at the start and end of the rallies instead of the average positions during the rallies. Furthermore, they divided the players into attacking and defensive groups, analysing the shape of the two groups separately. On top of that, they discovered that, given a dataset with formations of six teams, an unknown team formation could be correctly classified/assigned to one of the six teams. In basketball, Lucey *et al.* (2014) analysed defensive team formations of basketball players in the three seconds leading to a three-point shot attempt, finding they were able to predict whether the team was going to give up open shot opportunity or not.

At this point, we would like to stress the difference between team formation analysis, which is the topic of this chapter and is a spatial type of analysis, and the analysis to choose the optimal line-up of a team, which can benefit from player-specific data. The latter type of analysis focuses on the selection of actual players for each of the positions on the field and has been investigated more rigorously in, for example, hockey (Colleen Stuart, 2017), football (Barrick *et al.*, 1998; Tierney *et al.*, 2016), volleyball (Boon & Sierksma, 2003), basketball (Dezman *et al*., 2001) and cricket (Ahmed *et al.*, 2013).

8.1.2 Football

In football, teams generally aim to play according to a specific team formation (Kaminka *et al.*, 2003; Kuhlmann *et al.*, 2005), which can be defined as “A specific structure defining the distribution of players based on their positions within the field of play” (Ayanegui-Santiago, 2009). The first reports of different team formations in football date from 1963, not surprisingly the year that the offside rule was introduced in English football (Football bible, 2014). Ever since, new team formations have been introduced and changed, as football teams became more professional and teams started paying more attention to playing in a more organized manner. In football, team formations are generally labelled by a series of numbers connected by hyphens. Starting from the own goal and moving towards the opponents goal area, the number of players in each horizontal line are noted, thereby omitting the goalkeeper. The total number of players in the label of a team formation should therefore always sum up to ten, *i.e.* the number of field players. A 4-4-2, for example, means that there are four defenders, four midfielders and two attackers in the team formation.

Advantages of one specific team formation with respect to others, e.g. increased running distances when playing against a 4-2-3-1 instead of a 4-4-2, have been described by Carling (2011). Mapping the advantages of different team formations can be useful when comparing them with the own team strengths and weaknesses in order to choose the most suitable team formation for a game. Moreover, manipulating the positions off the players on the field has been found to be the key component of team tactics during football matches (Rein & Memmert, 2016). Team formation analysis in football can be performed in various ways, but often starts with a visual exploration and detection of formations in the football data (Stein *et al*., 2016; Stein *et al.*, 2018; Wu *et al.*, 2018).

Other methods focus on different Key Performance Indicators (KPIs), that are derived from the players' positions (Memmert *et al.,* 2017). For example, Sampaio & Macãs (2012) suggested the team centroid, team entropy, a team stretch index and the surface area of the team as key performance indicators for team formation analytics. Going further on this, Frencken *et al.* (2012) added the inter-team distance, *i.e.* the distance between the centroids of both teams, as a key performance indicator to detect goals or attempts in a match. Lemmink & Frencken (2013) demonstrated the possibility to use these key performance indicators not only for the entire team but also for subsets of the team such as players with specific roles, e.g. attackers or defenders. Reversing the approach, Memmert *et al.* (2019) later on investigated the effects of the use of different team formations on the above mentioned KPIs.

Bialkowski *et al.* (2014) proposed a method for the automatic detection of the type of team formation based on the average position of the players. They argue that, because the players swapping positions during the game, static ordering of the players does not accurately represent the team formation. In order to cope with this, they introduce dynamic ordering of players by the role that they occupy at a given instant in time. Using data from an entire Premiere League season, Lucey *et al.* (2013) and Bialkowski *et al*. (2014) found no significant difference between formations of different teams, but could detect that English Premier League teams used more offensive team formations during home games.

Various new methods use principles of (artificial) neural networks (McCulloch & Pitts, 1943). Visser *et al.* (2001) used artificial neural network systems to recognize the team formation of the opponent team. Starting with the positions of the opponent players at a certain timestamp, the neural network tried to classify that moment into a set of predefined team formations. Atmosukarto *et al.* (2013) later used an analogue method in American football and proposed the appropriate (most suitable) counter team formation for the own team. Going further on this work, Ayanegui-Santiago (2009) proposed to include multiple relations between players for the recognition of team formations in football. He divided the players into three groups (defenders, midfielders and attackers) and used labelled graphs between nodes of adjacent groups to describe and compare team formations.

The methods mentioned above generally aim at calculating frequencies of team formations. This facilitates comparison of different team formations and the temporal evolution of these team formations during the game (Grunz *et al.*, 2012). Furthermore, the occurrence of team formations can be linked to scoring goals and winning games, thus measuring the success of a specific team formation for a team. However, while most methods use quantitative metrics, Perin *et al.* (2013) argue that quantitative analysis is not sufficient to understand the team formation of a game or an entire season. Unfortunately, qualitative team formation analysis in football is at present mostly performed by human experts and is thus very labour intensive (Bialkowski *et al.*, 2014; Wu *et al.*, 2018). Moreover, few studies have analysed team formations on real size football fields, or on real football matches. Gonçalves *et al.* (2017), however, demonstrated that there are important differences between team tactics (and thus team formations) of matches played on a football field with a reduced size and games played on a full field. They therefore suggest that research should be performed on official football matches (meeting the official FIFA field size standards).

8.2 Datasets and analyses set-up

In this section, the datasets that are used in the football examples and the football experiment are introduced. After that, we present the analyses set-up of both the examples and the experiment, giving a detailed overview of the conducted research.

8.2.1 Datasets

Two datasets are analysed in this use case. The first dataset consists of 13 real football matches of a 2016-2017 professional football competition (of which one match was used for the use case presented in Chapter 6), and is used for the examples presented in Section 8.3. Due to privacy concerns, the teams and players are presented anonymously. During each of the 13 matches, all 22 players on the field were tracked by means of a camera system that was mounted in the stadium, logging their (x,y,t)-coordinates with a temporal resolution of 25Hz. Besides this, a manual annotation of the events occurring at each of the timestamps of each match is available (XML-file). The second dataset consists of two matches of the Belgian national football team at the 2018 FIFA World Cup in Russia and is used for the experiment presented in Section 8.4. This dataset contains the averaged positions of the 11 players of the Belgian national football team during the matches Belgium–Japan and Brazil-Belgium, which were extracted by manual logging from the official 'actual formation'-reports of the FIFA (2018). The 'actual formation'-reports are included in Appendix 5. Besides a theoretical team formation, filed by the team's coach in advance of the match, these reports contain the positions of the players averaged over every 15, 45 and 90 minutes of the match. The data mentioned in the reports takes the substitutions of players into account and displays a substituted player by means of a circle with a dashed line. Since each match report is publicly accessible (FIFA, 2018), no anonymous representation of this second dataset is required. For the performed analysis, the (x,y,t)-coordinates of the players are logged manually for each of the 18 actual performed averaged team formations (i.e. 9 for each match).

To evaluate the results of the examples and the experiment, it is necessary to compare the actual performed team formations with a ground-truth, *i.e.* a dataset containing the reference team formations and their common names in football. Unfortunately, no general accepted overview of reference team formations exists for football nor was it reported in literature (Zauli, 2003; Football bible, 2014; Wilson, 2014). Moreover, every coach or football analyst has his/her own interpretation of each team formation, based on experience or expectance of specific players roles for that formation (Wilson, 2014). There are, however, enormously popular football simulation games, such as the FIFA series that releases a new version every year (EaSports, 2019). To illustrate, more than 24 million copies were sold of its last years' version 'FIFA 18' (Hoggins, 2018). In fact, the series are so popular that many associate the word 'FIFA' with the football game rather than with the football federation. Up to 2016 (FIFA 16), the football game included a series of standard team formations. These formations are well-known by millions of people and are documented rigorously (FifaUteam, 2016), including a top-down view of the players positions for each of the standard formations. For the reasons mentioned above, we choose the standard formations of FIFA 16 as reference team formations for the use case presented in this chapter. The top-down view thereby allows for the (manual) logging of the coordinates of the players to create the reference team formations. An overview of the 25 reference team formations in football can be found in Appendix 5. Each of the reference team formations are named as mentioned in the first paragraph of Section 8.1.2.

8.2.2 Analysis set-up

In this use case, a series of real football examples and an experiment are used to demonstrate the use of the proposed method for team formation analysis (team tactics) in football. The approach differs from most of the research in this domain (see Section 8.1), where generally the aim is to study the effects of team formations on various football metrics. For this, team formations are mostly labelled by football experts, and usually no validation of these labels are provided. The aim of the proposed method, however, is to support the (automatic) detection (thus labelling) of team formations in football, based on a series of reference formations. In essence, it is aimed at supporting the football experts in their labelling of the team formations. A simple validation of this automated labelling can be done by a visual comparison of the detected label with the actual performed team formation. Further validation, however, is difficult and very time consuming, as no general accepted standard is available for this task. For the examples (Section 8.3) only the first type of validation is therefore provided. For the experiment, however, we provide a more thorough validation of the results, based on match reports published by popular sports media.

Only periods of the football matches where no players were sent off were selected for both examples as well as the experiment. Indeed, football players can be sent off during a football match by conceding one red card or two consecutive yellow cards. Such an event, however, would result in one team playing with less than 11 players and a QTC_S-matrix with different dimensions than the QTC_S-matrices that represent the reference formations (see Section 4.1). Consequently, it would not be possible to calculate distance between those QTC_S-matrices.

8.3 Football team formation analysis examples

In this section, a series of examples are given to demonstrate the use of the proposed method for team formation analysis in football. For this, the first dataset of Section 8.2.1 is used. In each of the examples, the viewpoint of football coach, analyst or reporter is chosen as a starting point. This is done to ensure that the examples are useful for providing real insight into the studied football matches (Memmert & Raabe, 2018). Consequently, the results are presented by means of similarity values instead of distance values (which were used in previous use cases), as these are more intuitive and thus easier to interpret for football experts with only a basic knowledge of the used method. This choice, however, does not influence the results, as the similarity values are the inverse of distance values (similarity = 1 – distance). The most basic version of QTC_S, *i.e.* level one QTC_S with two characters, is used for the examples. Furthermore, no static points are included. Considering the novelty of the method and the lack of a good ground truth (Feuerhake, 2016), the focus of this section primarily lies on the introduction of the possibilities rather than a validation of the results.

8.3.1 Compliance of the full team with desired team formation

Often, football coaches aim at using one or more predefined team formation(s) for their field players, according to the situation in the game (e.g. aggregate score, ball possession) and the team formation of their opponent. By using the proposed method to describe both the desired team formation(s) as well as the actual performed team formation, an evaluation of the compliance (similarity) of the team with the desired formation can be made, giving an indication of the team performance. For simplicity, we assume that, in this example, the individual players are assigned specific roles and a corresponding position on the field (e.g. a left-winger should position himself on the left attacking flank). The coach therefore does not allow players to switch roles and/or positions in the reference team formation. This implicates that permutations between players should not be taken into account when calculating distances between the QTC_S-matrices. This makes results easier to interpret, as this is the first example and the real guidelines of the coach are unknown.

The example based on the first dataset is shown in Figure 8-1. In this figure, the compliance of an anonymous team with a desired team formation is shown, during six different matches. In this example, the desired team formation is a variant of the 4-4-2 reference team formation (see Appendix 5), where the midfielders form a so-called 'diamond' on the midfield (DiBernardo, 2014), which is shown on the right of Figure 8-1. The analysis is performed on the positions of the players which were averaged for every 5 minutes of the match, excluding the goalkeeper and without player permutations. The higher the similarity in the graph, the more the actual team formation resembled the desired 4-4-2 (with diamond on the midfield), during the game. It can be seen that during match 1 the compliance of the team is the highest, while match 6 is characterized by a low compliance with the desired team formation. In general, the similarity of the actual performed team formation with the desired team formation are relatively low. This is probably because permutations between players are not taken into account in this example, while role switches of players are reported to occur frequently in real football matches (Bialkowski *et al.*, 2014).

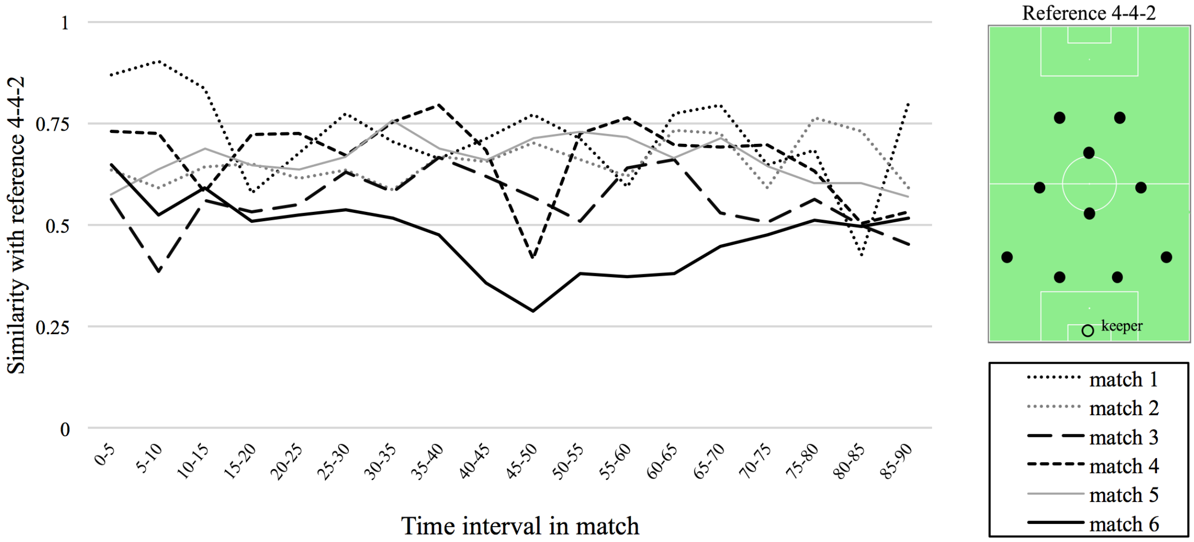


Fig. 8-1. Similarity of an anonymous team (excluding the goalkeeper) with a desired 4-4-2 formation during 6 matches, with a temporal resolution of 5 minutes and without permutations.

8.3.2 Analysis of parts of the team formation

Going more in detail, it can be interesting to analyse how different groups of players of a team, *e.g.* defenders and midfielders, each stick to their desired formation during a match. Figure 8-2 displays the compliance of the midfielders and defenders of an anonymous team with their respective reference formations throughout one match. The analysis is performed using a temporal resolution of 5 minutes and without taking player permutations into account. From the graph it can be seen that the defence much more sticks to its desired formation throughout the game than the midfield, which naturally has a more flexible and interchanging character (Gonçalves *et al.*, 2014). Between minutes 15 and 30 of the match, however, the only period during which the displayed team conceded (multiple) goals, both defenders as well as midfielders had the highest deformations with respect to their reference formations. Again, the relatively low similarities of the midfield can be assumed to be because by players switching positions (Bialkowski *et al.*, 2014).

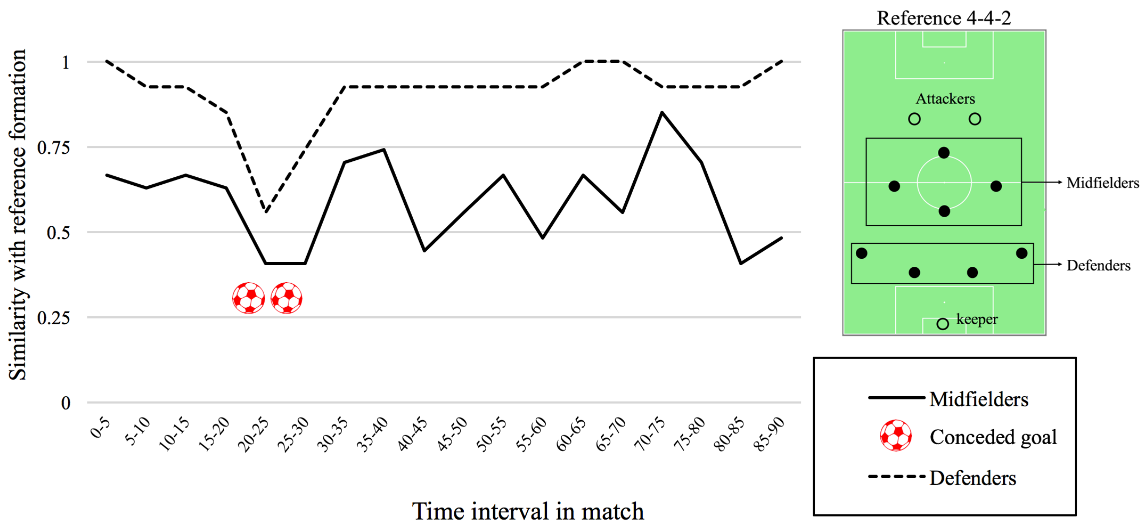


Fig. 8-2. The similarity of midfielders and defenders of an anonymous team with their desired formation throughout one game, with a temporal resolution of 5 minutes and without permutations.

8.3.3 Analysis of a teams playing style

In the two previous sections, one desired team formation is used to evaluate the actual performed team formation. If a coach or football expert would want to detect the occurrence a series of reference team formations during one or more matches, the actual team formation can be compared with all of the 25 reference team formations (see Section 8.2.1). This can be seen as an attempt of trying to define the playing style of football team. Indeed, Fernandez-Navarro *et al.* (2016) argue that the playing style of a team can be defined as the general behaviour of a group of players or whole team, that is aimed at achieving the defensive and offensive team objectives in a match. Of this behaviour, the evolution of the positions of the players on the field to execute the team tactics is one of the most important aspects (Rein & Memmert, 2016). An example of this is shown in Figure 8-3, where for two matches of an anonymous team, frequencies of the most similar reference team formation at every second of the game are displayed. This analysis excludes the goalkeeper and does not take player permutations into account. The results illustrate the variety of team formations that are performed by one team during a game or even between different games. However, if permutations would be taken into account, a drop of the absolute number of detected reference formations is to be expected. As such, a team's playing style, *i.e.* a set of regular played team formations by a team, can be defined and compared between teams and matches by using the proposed method for static sports analytics. It might be useful for a football coach, analyst or reporter to analyse the playing style of an opponent team in such a manner for the preparation of a real match, assuming he/she has access to the necessary positional data.

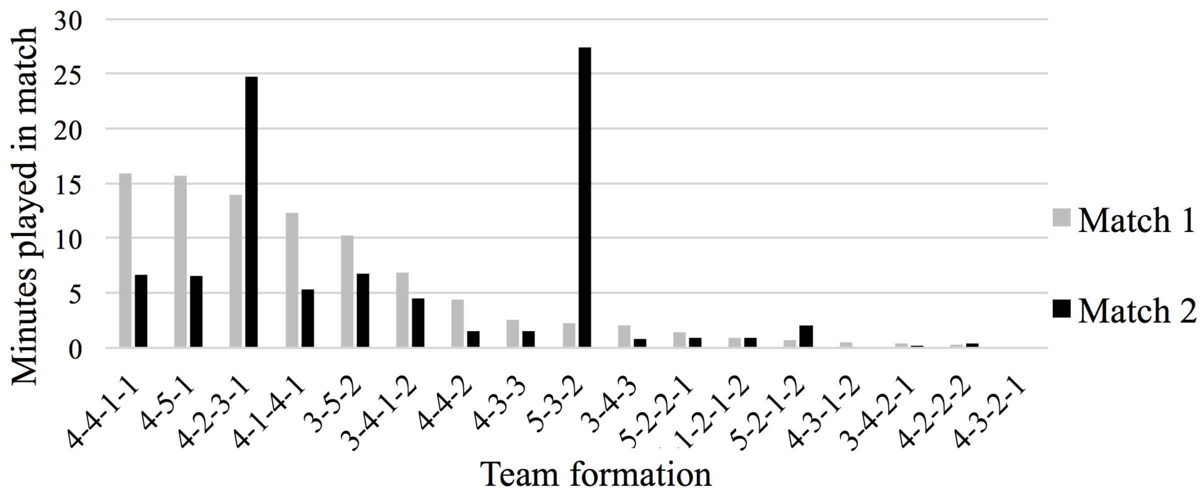


Fig. 8-3. Frequency of the most similar reference team formations played by an anonymous team during two matches, ordered according to the frequencies of match 1. The analysis is performed with a temporal resolution of 1 second and player permutations are not taken into account.

8.4 Experiment

In this experiment, we aim at analysing the team formation of the Belgian national football team during two matches of the 2018 FIFA World Cup. More specifically, we try to automatically detect according to which reference team formation the team is playing during the matches Belgium-Japan and Brazil-Belgium. To that end, we make use of the second dataset that was introduced in Section 8.2.1. As for these two matches (official) match reports are available, the results of the automatic detection of the team formation can be validated. Furthermore, we include an additional validation of the results based on match reports published by popular sports media.

For the analysis, different set-ups are possible. Iterative preliminary tests were conducted to decide the best configuration of the parameters for the experiment (see Section 5.1.2). Static points, for example, could be used to fix the rotation of the team formation or, in combination with level two QTC_S, to take the scale into account. For this experiment, however, problems with rotation are not to be expected as football players naturally try to play towards the opponents' goal. In fact, the goalkeeper of the team can almost be seen as a static point. Indeed, the goalkeeper will always be positioned in front of his own goal, thus with an almost identical orientation with respect to the rest of the players of his team. Given the qualitative nature of QTC, small positional changes of the goalkeeper's position do not pose a problem in this respect. Consequently, all 11 players were included in the analysis, but without static points. Including 11 players in the analysis implies that the QTC_S-matrix has 110 rows and 110 columns, thus containing 5995 cells that each store a QTC-relation. Given these large dimensions, the QTC_S-variant with the lowest level of detail (level one QTC_S) is used. In this way, a detailed description of the team formation is guaranteed while the risk of over-averaging important changes in the positioning of individual players is reduced, when compared with the more detailed level two QTC_S. For reasons of simplicity, no player permutations were included in the examples presented in Section 8.3. However, when the coach's intentions are not clear, it is in fact better to take these permutations into account when calculating distances between QTC_S-matrices. Because all 11 players are included in the analysis, including all the permutations would mean that 39916800 permutations of the QTC_S-representation of each actual performed team formation would have to be compared with each of the 25 reference team formations. Given the limitations of the current implementation of the static sports analytics method (see Section 5.1.4), it is not possible to calculate such a high number of (large) QTC_S-matrices nor to compare them with the QTC_S-representations of the reference team formations. As an alternative, 4 of the 11 players are chosen as 'fixed' based upon common sense, thereby reducing the number of players to permute to 7. Assuming that a football team always plays with one goalkeeper and at least three defenders, these players can be omitted from the list of players for which permutations have to be calculated. Indeed, in the reference team formations there are no team formations without goalkeeper or with less than 3 defenders. Furthermore, defensive players are characterized by a high level of positional rigidity (Gonçalves *et al.*, 2014), meaning that position switches are not likely to occur. This allows to reduce the number of permutations to 5040, which is possible to calculate with the current implementation.

To summarize, level one QTC_S is used to analyse the team formation of the 11 players of the Belgian national football team during two matches, without static points and with permutations for 7 of the 11 players (excluding the goalkeeper and three defenders). As was done for the examples of Section 8.3, similarities are reported instead of distances. Eventual substitutions are automatically accounted for in the dataset (see Appendix 5).

8.4.1 Results

Figure 8-4 displays the results of the experiment for the matches Belgium-Japan and Brazil-Belgium. For each of the actual performed team formations included in the official FIFA match reports (9 per match), the most similar reference team formation is shown (column 'label') along with the similarity (1- distance) with that reference team formation (column 'similarity with label'). Besides the actual performed team formations, the figure includes the theoretical team formations that were submitted by the team's coach to the FIFA before the start of both matches. While the theoretical team formation is a 3-4-3 for both matches, this formation is never detected as an actual performed team formation. Instead, small variants of this team formation are reported to have occurred during the matches. In the match against Japan, for example, the highly similar 3-4-2-1 formation was played by the team for all of the considered time intervals. In fact, the 3-4-2-1 formation can be seen as a variant of the more general 3-4-3, with the only difference being that the two flank midfielders play a bit higher (*i.e.* more towards the opponent's goal) and the two flank attackers play a bit closer to the middle attacker (see Figure 8-5). In the match against Brazil, the team most commonly played according to a 3-4-1-2 formation which is another variant of the 4-3-3 formation. In this formation, the flank midfielders play a bit higher and the flank attackers play a bit closer to the centre of the field while the middle attacker plays significantly lower. In fact, the latter is the only important difference between the 3-4-1-2 and the 3-4-2-1 formations. Both variants, however, are very similar to the 3-4-3 formation as can be seen from Figure 8-5. More surprisingly, there are two time intervals (minutes 15-30 and 75-90) during the match against Brazil for which reference team formations with four defenders are detected as most similar (shown in yellow in Figure 8-4). This is surprising, as the theoretical team formation only shows three defenders.

In general, the reported similarities of the actual team formations with the labels are higher than the similarities reported in the examples of Section 8.3. There are two reasons for this. The first is that the label is defined as the reference team formation with the maximum similarity for the considered actual team formation, while for the examples (see Figures 8-1 and 8-2) similarities with just one reference team formation were calculated and reported. The second reason is that player permutations are taken into account. For each actual team formation, 7 of the players are swapped places consecutively and for every configuration, the similarity with each of the reference team formations is calculated. For each reference team formation, the maximum similarity is saved. From those similarities, only the team reference formation with the maximum similarity is withheld and displayed in Figure 8-4. From this overview it can be seen that, in general, the similarities of the labels are in general lower in the second match.

	ACTUAL TEAM FORMATION	LABEL	SIMILARITY WITH LABEL
BELGIUM-JAPAN	0-15	3-4-2-1	0.9341
	15-30	3-4-2-1	0.9483
	30-45	3-4-2-1	0.9491
	0-45	3-4-2-1	0.9508
	45-60	3-4-2-1	0.8657
	60-75	3-4-2-1	0.9358
	75-90	3-4-2-1	0.9341
	45-90	3-4-2-1	0.9341
	0-90	3-4-2-1	0.9458
	Theoretical formation	3-4-3	0.9383
BRAZIL-BELGIUM	0-15	3-4-1-2	0.8882
	15-30	4-4-1-1	0.9066
	30-45	3-4-1-2	0.9141
	0-45	3-4-1-2	0.9158
	45-60	3-4-1-2	0.8924
	60-75	3-4-1-2	0.9024
	75-90	4-1-2-1-2	0.8974
	45-90	3-4-1-2	0.9016
	0-90	3-4-1-2	0.9091
	Theoretical formation	3-4-3	0.9383

Fig. 8-4. Results of the experiment showing the reference team formation with the highest similarity for each of the actual performed team formations of the Belgian national football team during two matches of the 2018 FIFA World Cup.

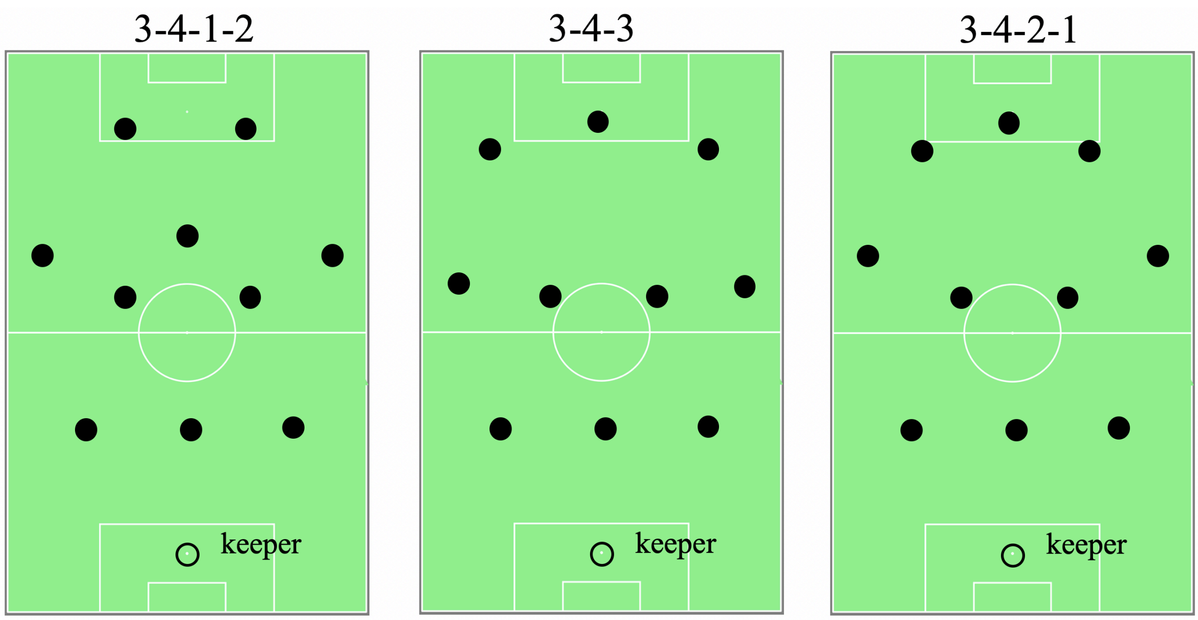


Fig. 8-5. The more general 3-4-3 reference team formation and two variants played by the Belgian national football team.

8.4.2 Validation

A first validation of the results is the visual comparison of the actual performed team formations with the reference team formations. Indeed, when making a visual comparison of the actual team formations with the different reference team formations, we are convinced that (roughly) the correct labels were assigned to the respective time intervals. Consequently, the results (*i.e.* the labels in Figure 8-4) can be compared with the theoretical team formations that were filed by the team's coach before the start of each match. The coach filed a 3-4-3 formation for both matches, which is very similar to the 3-4-3 variants that were detected for the whole match against Japan (3-4-2-1) and for the most part of the match against Brazil (3-4-1-2). This can be seen from the hierarchical clustering (UPGMA) of all the reference team formations shown in Figure 8-6. This hierarchical clustering was created using an identical set-up, with the parameters set according to the description in the first two paragraphs of Section 8.4. The clustering indeed shows the similarities between the theoretical 3-4-3 and the detected variants. As the team is not obliged to play according to the theoretical team formation, this finding cannot be seen as conclusive evidence. It is, however, valuable as an early step of the validation and invites to further validate the results.

For a more thorough validation, the results can be compared with the match reports that were published by football experts in popular sports media. In general, the Belgian national team has reported to play according to a 3-4-3 formation during the 2018 FIFA World Cup:

> *"Belgium, during qualifying for this tournament and at the World Cup itself, have played almost exclusively in an adventurous 3-4-3 formation that packs the pitch with attackers and seeks to dominate the ball."* – Deadspin (Haisley, 2018)

> *"And unlike the great, gritty 1980s' Devils, this vintage plays gorgeous football. ….because Martinez has fielded all his creators, in a 3-4-3 formation ..'"* – Financial Times (Kuper, 2018)

For the match against Japan, several popular sports media reported a 3-4-2-1 formation of the Belgian national football team (ESPN, 2018; Transfermarkt, 2018), similar to the results of the experiment (see Figure 8-4). Other media, however, reported that the Belgian national football team played according to the more general 3-4-3 formation during the match against Japan:

> "*Belgium started in a 3-4-3 and many senior players returned to the starting line-up from their game against England. Japan lined up in a 4-2-3-1.*" – Business Standard (2018)

However, this match report includes a visual overview of what they refer to as the 3-4-3 formation (Business Standard, 2018). From this visual overview it can be seen that this team formation much more resembles the 3-4-2-1 reference team formation than the 3-4-3 reference team formation that is used in this use case (see Figure 8-5). Indeed, from the visual overview it can be seen that the flank midfielders play higher and the flank attackers closer to each other than in the 3-4-3 reference team formation. For this reason it can be assumed that in their tactical analysis, the analysts do not distinguish between the different variants of the 3-4-3 formation shown in Figure 8-5. Their report can therefore be seen as a positive validation of the results shown in Figure 8-4, especially considering the immense diversity of reference team formations that could be assigned as label (see Appendix 5).

For the match against Brazil, various popular sports media reported that the Belgian national football team alternated between a defence with three defenders (in ball possession) and four defenders (when defending), in agreement with the results of the experiment. Indeed, during the match against Brazil the team formation of the Belgian national football team differed quite a lot from the team formations of their previous matches:

> "*On one hand, Belgium coach Roberto Martinez stunned Brazil by using an entirely unexpected system, deploying key players in new roles and shifting smoothly*

between a three-man defence and a back four." – Independent Newspaper (Cox, 2018)

"*Head coach Roberto Martinez brought Marouane Fellaini and Nacer Chadli into Belgium's starting line-up, with Dries Mertens and Yannick Carrasco dropping to the bench. They continued with their defensive back three when in possession, but often converted to a back four when defending.*" – The Coaches Voice (2018)

Other reports go beyond describing the changing number of defenders and describe the adjustments made to the team formation, thus including all the players:

"*As can be seen above, this structure occurred when De Bruyne dropped deep to create a 4-3-1-2 deep block in order to fashion a 4v3 against Neymar, Coutinho, and Marcelo.*" – Outside of The Boot (Arvind, 2018)

Unfortunately, the 4-3-1-2 formation mentioned by Arvind (2018) is not included in the results shown in Figure 8-4. The 4-1-2-1-2 formation which is detected as label for the time interval between the 75th and the 90th minute, however, is very similar to the reported 4-3-1-2 formation, as can be seen from the hierarchical clustering of the reference team formations shown in Figure 8-6. Looking at the time interval between the 15th and the 30th minute, the 4-4-1-1 formation label cannot be explained by using the hierarchical cluster tree. Indeed, the 4-4-1-1 formation seems to have rather large distances with the expected formations (*i.e.* the 4-3-1-2 formation). Remarkably, the similarity with the label is rather low for this time interval (15th -30th minute) when compared to the other labels of this match and especially with the labels of the match against Japan. This means that the Belgian national football team did not play strictly according to one of the reference team formations. The importance of the label for this time interval should therefore not be overestimated. Given the low similarity, it can be interesting to take a look at the similarities with the other reference team formations for that time interval. Figure 8-7 displays the ten reference team formations that have the highest similarities with the actual performed team formation for the time interval between the 15th and the 30th minute for the match against Brazil. From this overview it can be seen that most of the other expected formations with 4 defenders are ranked among the most similar reference team formations, such as the various variants of the 4-3-3 formation (Smith, 2018). Furthermore, it can be seen that a large number of reference team formations have relatively similar similarity values with respect to the actual performed team formation. We can therefore conclude that during this time interval, the team did not play strictly according to one of the reference team formations.

The strict definition of the time interval can be one of the reasons for this, as during a rather long period of 15 minutes a team can play according to multiple actual performed team formations that better match with one of the reference team formations. During these longer time intervals, however, the players coordinates are averaged, meaning that these labels with higher similarities might not be detected.

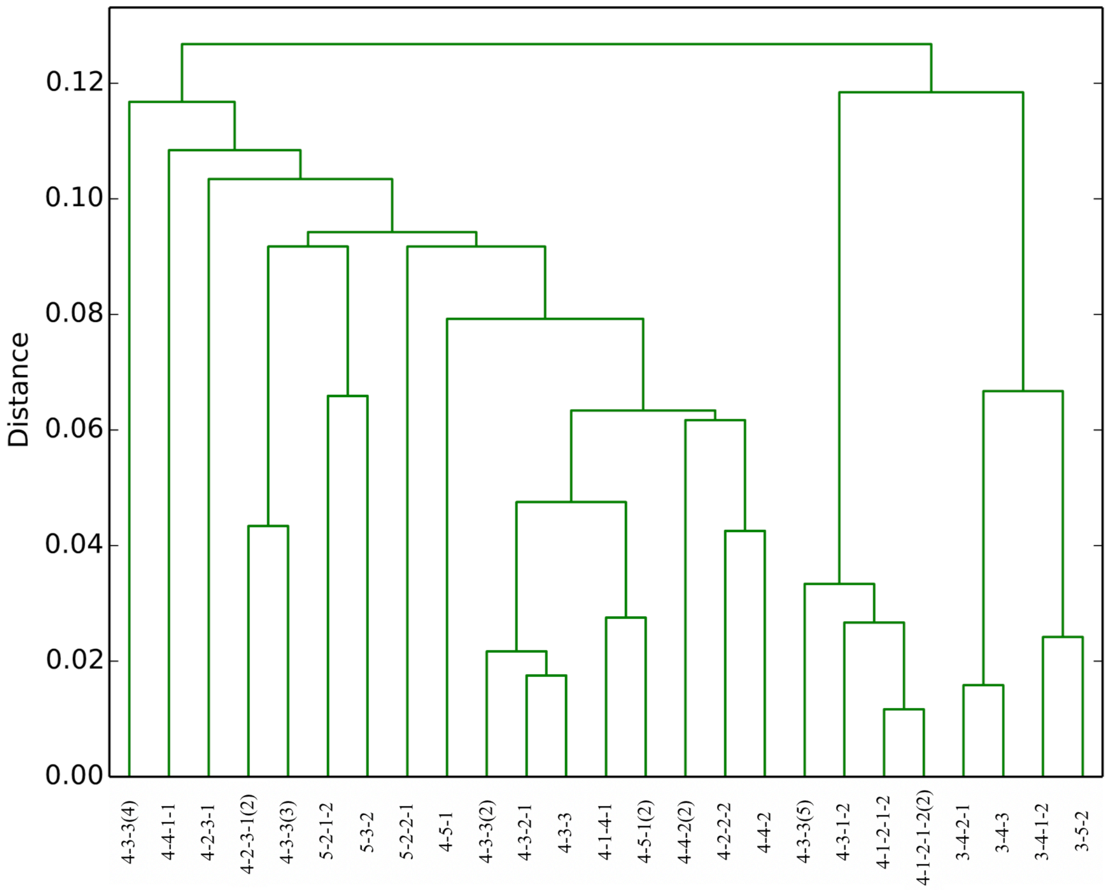


Fig. 8-6. Hierarchical clustering (UPGMA) of the reference team formations using the parameters of the experiment.

BELGIUM - BRAZIL 15-30		
RANK	REFERENCE TEAM FORMATION	SIMILARITY
1	4-4-1-1	0.9066
2	3-4-1-2	0.8907
3	4-3-3(3)	0.8866
4	4-2-3-1(2)	0.8866
5	5-3-2	0.8816
6	4-5-1	0.8816
7	3-5-2	0.8816
8	3-4-3	0.8816
9	4-3-2-1	0.8791
10	4-3-3(2)	0.8782

Fig. 8-7. The ten highest ranked reference team formations of the actual performed team formation of the Belgium national team between the 15th and the 30th minute during the match against Brazil.

8.5 Discussion

In this section, the importance of the results for the domain of team formation analysis in football are discussed. As shown by the different examples and the experiment, the proposed method for static sports analytics allows for the analysis of team formations in football. In this section we discuss the results and present some of the advantages and limitations of the proposed method for team formation analysis in football. An evaluation of the use case for the sports analytics domain and its value for answering the respective research questions can be found in Chapter 9.

An important advantage of the proposed method is that players can be allowed to take each other's place in a team formation at different moments. This can be done by allowing permutations (all of them or only specific ones) while calculating the distance between different QTC_S-matrices. Because of this characteristic, the method takes more into account than just the static ordering of the players on the field, which was argued to be a prerequisite for a successful team formation analysis in football by Bialkowski *et al.* (2014). Unlike most other established methods (e.g. the methods introduced by Ayanegui-Santiago (2009) and Visser *et al.* (2001)), the proposed method might therefore give better results when specific players switch positions and/or roles during the game, as is often reported for so-called utility players (Schuth *et al.*, 2015). Although all of the players can theoretically change positions and/or roles during the game, position and role switches most often occur between flank players and attackers (Schuth *et al.*, 2015). As position and/or role switches can be desired by a team's coach, it is possible to allow these by only calculating only specific permutations of the QTC_S-matrices during the distance calculation.

Another advantage is the wide applicability of the static sports analytics method in sports. Besides analysing player positions that are averaged over longer periods, it allows for the analysis of team formations at set pieces. Set pieces in football are moments where the ball is returned to open play, *i.e.* corners kicks, free kicks or throw-ins. Indeed, set pieces are getting increasingly important in modern football as opportunities of scoring a goal (Yiannakos & Aramats, 2004). During the 2018 FIFA World Cup, for example, 73 of the in total 169 scored goals (43.2%) resulted from set pieces (excluding throw-ins). Furthermore, teams have reported (Alcock, 2010) to spend more time on the training of set pieces, resulting in an increasing efficiency, *i.e.* a higher ratio of the number of scored goals resulting from a set piece divided by the number of set pieces. A famous example is the Danish club Midtylland who, through extensive statistical analysis and positional optimization of their players at set pieces, achieved to score a league record of 25 goals from set pieces in the 2014-2015 season of the Danish first league (Ingle, 2015). In general, set pieces occur regularly during football matches, with the average number of corner kicks per match ranging from 10.2 to 10.9 per match (De Baranda & Lopez-Riquelme, 2012; Wallace & Norton, 2014). However, not all types of set pieces are occurring in equal numbers during the game, and their effectiveness for scoring goals differs significantly depending on the type. Throw-ins, for example, have reported to be the least effective while free kicks have proven the most effective for scoring goals (Carling *et al.*, 2005). For free kicks, the effectiveness is largely defined by their location on the field. For these reasons, it can be interesting to study the team formation of both the attacking as well as the defending team at the moment of the set pieces. That the team formation at set pieces is increasingly used by coaches to execute team tactics can be understood from the research of Wallace & Norton (2014). They demonstrated that, in order to achieve a perfect attacking team formation, the preparation time of the offensive team for a direct free kick increased significantly in recent years. Moreover, based on FIFA World Cup data from 1966 to 2010, the preparation time of free kicks that resulted in a direct shot at goal have almost doubled (Wallace & Norton, 2014). Indeed, team formations at set pieces are one of the most studied and trained aspects of team formation in football (Sarmento *et al.,* 2014). By transforming both the desired as well as the actual performed formations at set pieces into QTC_S-matrices and comparing them, coaches could get an overview of whether and to what extent the ideal trained formation of their own team was achieved in real matches or could get insight into the tactics and regularly performed team formations at set pieces of opponent teams.

One of the challenges for the analysis of set pieces with the proposed method is the definition of the players involved in the set piece. Not all of the 11 players (or 22, if both teams are taken into account) can be assumed to actively participate in the attack or defence directly following a set piece. It is therefore advisable that a football expert (e.g. coach) defines which of the players should be taken into account at each set piece separately. When multiple set pieces need to be compared, the number of involved players might therefore differ across these set pieces. Consequently, the dimensions (and labels) of the QTC_S matrices will differ, making it impossible to calculate a distance between them with the current implementation of the method (see Section 4.1).

A disadvantage is that when a player is substituted during a match, the team formation analysis cannot be performed for a period containing the moment of this substitution. As the substituted player goes out of the field, his/her average position is influenced by this displacement that should in fact be neglected as it has no influence on the game. The largest problem, however, is that this player might be replaced with another player that will play at a different position. When the substituted player is replaced by a player that plays at the same position and/or has a similar role, the average positions are often reported without interruption (as is the case for the actual formations reports of the 2018 FIFA World Cup, shown in Appendix 5). In that case, only the label of the player is adjusted when a substitution occurs during the period over which the average positions are calculated. Since during one game three substitutions are allowed for each team, it is advisable to apply this approach in order to be able to use the proposed method for the analysis of team formations in football.

Another disadvantage of the proposed method for team formation analysis in football is the large number of QTC-relations (*i.e.* cells) in one QTC_S-matrix. Indeed, even when the formation of a limited number of players is described, the QTC_S-matrix has relatively large dimensions (see Section 4.3). This is because all the vectors between all of the players are described with respect to each other, by means of a QTC-character in the QTC_S-matrix. The advantage is that the formation is described with enough detail, even when only a limited number of players are analysed. The downside, however, is that there is a risk of over-averaging small but significant differences in the positioning of individual players at different moments when calculating the distance between two QTC_S-matrices. The reason for this is that all of the differences between two team formations are summarized in one distance value. To address this, the number of vectors that are included in the matrix could be reduced, e.g. by only including the vectors that describe the convex hull of the team formation at that moment, or by excluding 'double' vectors. Double vectors are vectors that are constructed between two identical players, but have opposite directions. By omitting one of a pair of double vectors, the number of vectors can be reduced by 50%, causing an linear (quadratic) decrease in the number of cells in the respective QTC_S-matrix (see the formula in Section 4.3). Similarly, Ayanegui-Santiago (2009) only takes relations between players of direct adjacent groups (defenders, midfielders and attackers) into account. An alternative way of increasing the impact of the position of certain players on the resulting distance value is by simply assigning more weight to the vectors that have the player as start or endpoint in the distance calculation. The downsides of this procedure is that it might not be easy to find suitable weights and that it requires an additional effort of a user, reducing the advantages of the automated analysis.

When analysing the positions of football players averaged over longer time intervals rather than positions at exact moments (e.g. at set pieces), it is important to realise that some problems might occur while calculating the averaged player coordinates.

As mentioned earlier, a time interval can include one or multiple moments where a player is substituted, making the analysis of the team formation for that interval problematic. Similarly, when a player is sent off the field, it is not possible to compare the team formation of time intervals before that event with the team formations of intervals after that event using the proposed method, because the QTC_S-matrices will have different dimensions. There are many other events in football that can influence the reliability of coordinates that are averaged over time. A time interval can, for example, include one or multiple set pieces which require a rather long preparation time, which has proven to be gradually increasing in recent years (Wallace & Norton, 2014). During the preparation time of a set piece, the players do not occupy their normal positions on the field, which has a negative impact on the reliability of the averaged player positions for reflecting the team formation. Similar problems arise when a player is injured and treated on the field or when players run towards the referee to contest a decision during the match.

Indeed, as mentioned in the validation section of the experiment (Section 8.4.2), an important challenge is the correct definition of the time intervals for which team formations are studied. When a time interval is too long, it might include multiple periods during which the team deliberately played according to a different reference team formation. As the coach usually defines the requested team formation at each moment, only (s)he can correctly define the intervals to study. Football analysts, on the other hand, can make use of match reports to try to find suitable time intervals to analyse team formations. Indeed, requested team formations often change after substitutions or when a goal is scored. Nevertheless, the definition of the time intervals remains a time-consuming effort for a football expert. In the experiment this was bypassed by using the time intervals provided in the official FIFA match reports. The downside, however, is that the time intervals could not be corrected for the events influencing the averaged positions mentioned in the previous paragraph. This might be the reason why for some of the time intervals labels with relatively low similarity were found. With respect to this, it is important to note that when similarities with the labels are low, it is hard to draw conclusions from the 'detected' reference team formation (label).

REFERENCES

Ahmed, F., Deb, K., Jindal, A. (2013). Multi-objective optimization and decision making approaches to cricket team Selection. *Applied Soft Computing*, 13(1), 402-414.

Alcock, A. (2010). Analysis of direct free kicks in the women's football World Cup 2007. *European Journal of Sport Science*, 10, 279-284.

Arvind, O. (2018). Tactical analysis: Brazil 1-2 Belgium/ Belgium nick a fortunate win through decisive counter attacks. [online] [viewed 19/04/2019]. Available from: http://outsideoftheboot.com/2018/07/08/tactical-analysis-brazil-1-2-belgium-belgium-nick-a-fortunate-win-through-decisive-counter-attacks

Atmosukarto, I., Ghanem, B., Ahuja, S., Muthuswamy, K., Ahuja, N. (2013). Automatic recognition of offensive team formation in American Football plays. Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (pp. 991-998). Portland, OR, U.S.A.

Ayanegui-Santiago, H. (2009). Recognizing Team Formations in Multiagent Systems: Applications in Robotic Soccer. In N.T. Nguyen, R. Kowalczyk, S.M. Chen (Eds.), Computational Collective Intelligence. Semantic Web, Social Networks and Multiagent Systems. Lecture Notes in Computer Science (pp. 163-173). Berlin, Heidelberg, Germany: Springer.

Barrick, M.R., Stewart, G.L., Neubert, M.J. (1998). Mount relating member ability and personality to work-team processes and team effectiveness. *Journal of Applied Psychology*, 83, 377-391.

Bialkowski, A., Lucey, P., Carr, P., Yue, Y., Matthews, I. (2014). "Win at Home and Draw Away": Automatic Formation Analysis Highlighting the Differences in Home and Away Team Behaviors. *Disney Research.*

Boon, B.H., Sierksma, G. (2003). Team formation: Matching quality supply and quality demand. *European Journal of Operational Research,* 148(2), 277-292.

Business Standard. (2018). In a great footballing spectacle, Belgium win match, Japan hearts. FT 3-2. [online] [viewed 25/04/2019]. Available from: https://www.business-standard.com/article/sports/fifa-world-cup-2018-belgium-vs-japan-live-score-updates-hazard-lukaku-kagawa-play-round-of-16-watch-free-live-streaming-fixtures-schedule-of-world-cup-in-russia-118070200689_1.html

Carling, C. (2011). Influence of opposition team formation on physical and skill-related performance in a professional soccer team. *European Journal of Sport Science*, 11(3), 155-164.

Carling, C., Williams, A.M., Reilly, T. (2005). Handbook of soccer match analysis: a systematic approach to improving importance. London, U.K.: Routledge.

Colleen Stuart, H. (2017). Structural disruption, experimentation and performance in professional hockey teams: A network perspective on member change. *Organization Science*, 28(2), 283-300.

Cox, M. (2018). World Cup 2018: Belgium out-tactic Brazil to reach semi-finals as Roberto Martinez channels his channels his Everton days. [online] [viewed 19/04/2019]. Available from: https://www.independent.co.uk/sport/football/world-cup/world-cup-2018-belgium-vs-brazil-tactics-roberto-martin-everton-romelu-lukaku-eden-hazard-a8436056.html

De Baranda, P.S., Lopez-Riquelme, D. (2012). Analysis of corner kicks in relation to match status in the 2006 World Cup. *European Journal of Sport Science*, 12(2), 121-219.

Dezman, B., Trinic, S., Dizdar, D. (2001). Expert model of decision making system for efficient orientation of basketball players to positions and roles in the game - empirical verification. *Collegium Antropologicum,* 25(1), 141-152.

DiBernardo, M. (2014). Coaching the modern 4-4-2 diamond soccer formation: Tactics & training exercises. Unknown Publisher.

EaSports. (2019). FIFA 2019. [online] [viewed 12/03/2019]. Available from: https://www.easports.com/fifa

ESPN. (2018). Chadli late show sees Belgium past Japan. [online] [viewed 23/04/2019]. Available from: http://www.espn.com/soccer/match?gameId=498149

Fernandez-Navarro, J., Fradua, L., Zubillaga, A., Ford, P.R., McRobert, A.P. (2016). Attacking and defensive styles of play in soccer: analysis of Spanish and English elite teams. *Journal of Sports Sciences*, 34(24), 2195-2204.

Feuerhake, U. (2016). Recognition of Repetitive movement patterns-the case of football analysis. *International Journal of Geo-Information*, 5(11), 208.

FIFA. (2018). 2018 FIFA WORLD CUP RUSSIA. [online]. [viewed 14/12/2018]. Available from: https://www.fifa.com/worldcup/matches/

FifaUTeam. (2016). FIFA 16 ultimate team formations guide. [online] [viewed 12/04/2018]. Available from: https://www.fifauteam.com/fifa-16-ultimate-team-formations/

Football bible. (2014). Old football formations. [online] [viewed 14/03/2019]. Available from: https://www.football-bible.com/soccer-info/old-football-formations.html

Frencken, W., de Poel, H., Visscher C., Lemmink, K. (2012). Variability of inter-team distances associated with match events in elite standard-soccer. *Journal of Sport Sciences*, 30(12), 1207-1213.

Gonçalves, B., Esteves, P., Folgado, H., Ric, A., Torrents, C., Sampaio, J. (2017). Effects of pitch area-restrictions on tactical behaviour, physical, and psychological performances in soccer large-sided games. *The Journal of Strength & Conditioning Research*, 31(9), 2398-2408.

Gonçalves, B., Figueira, B.E., Maçãs, V., Sampaio, J. (2014). Effect of player position on movement behaviour, physical and physiological performances during an 11-a-side football game. *Journal of Sports Sciences*, 32(2), 191-199.

Grunz, A., Memmert, D., Perl, J. (2012). Tactical pattern recognition in soccer games by means of special self-organizing maps. *Human Movement Science,* 31(2), 334-343.

Hailsey, B. (2018). Belgium and Brazil just put on the best show of the World Cup. [online] [viewed 23/04/2019] Available from: https://deadspin.com/belgium-and-brazil-just-put-on-the-best-show-of-the-wor-1827406328

Hoggins, T. (2018). Electronic Arts eyes revenue boost with upcoming FIFA 19 release. [online] [viewed 13/03/2019]. Available from: https://www.telegraph.co.uk/gaming/news/electronic-arts-eyes-revenue-boost-upcoming-fifa-19-release/

Ingle, S. (2015). How Midtjylland took the analytical route towards the champions league. [online] [viewed 10/12/2018]. Available from: https://www.theguardian.com/football/2015/jul/27/how-fc-midtjylland-analytical-route-champions-league-brentford-matthew-benham

Jäger, M.J., Schöllhorn, W.I. (2012). Identifying individuality and variability in team tactics by means of statistical shape analysis and multilayer perceptrons. *Human Movement Science,* 31(2), 303-317.

Kaminka, G.A., Fidanboylu, M., Chang, A., Veloso, M.M. (2003). Learning the sequential coordinated behavior of teams from observations. In G.A. Kaminka, P.U. Lima, R. Rojas, (Eds.). RoboCup 2002: Robotic soccer world cup VI. Lecture Notes in Computer Science (pp. 111-125). Berlin, Heidelberg, Germany: Springer.

Kuhlmann, G., Stone, P., Lallinger, J. (2005). The UT Austin villa 2003 champion simulator coach: A machine learning approach. In D. Nardi, M. Riedmiller, C. Sammut, J. Santos-Victor (Eds.). RoboCup 2004: Robot Soccer World Cup VIII. Lecture Notes in Computer Science (pp. 636-644). Berlin, Heidelberg, Germany: Springer.

Kuper, S. (2018). Adaptable Red Devils give Belgium best shot at winning the World Cup. [online] [viewed 25/04/2019]. Available from: https://www.ft.com/content/70abbff8-7d37-11e8-bc55-50daf11b720d

Lemmink, K., Frencken, W. (2013). Tactical performance analysis in invasion games: Perspectives from a dynamical system approach with examples from soccer. In T. McGarry, P. O'Dono-ghue, J. Sampaio (Eds.). Routledge Handbook of Sports Performance Analysis (pp. 89-100). London, U.K.: Routledge.

Lucey, P., Bialkowski, A., Carr, P., Morgan, S., Matthews, I., Sheikh, Y. (2013). Representing and discovering adversarial team behaviors using player roles. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (pp. 2706-2713).

Lucey, P., Bialkowski, A., Carr, P., Yue, Y., Matthews, I. (2014). "How to get an open shot": Analyzing team movement in basketball using tracking data. *Disney research*.

McCulloch, W.S., Pitts, W. (1943). A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5(4), 115-133.

Memmert, D., Lemmink, K.A.P.M., Sampaio, J. (2017). Current approaches to tactical performance analysis in soccer. *Sports Medicine*, 47(1), 1-10.

Memmert, D., Raabe, D. (2018). Data Analytics in Football: Positional Data Collection, Modelling and Analysis. London, U.K.: Routledge.

Memmert, D., Raabe, D., Schwab, S., Rein, R. (2019). A tactical comparison of the 4-2-3-1 and 3-5-2 formation in soccer: a theory-oriented, experimental approach based on positional data in an 11 vs 11 game set-up. *PLOS One*, 14(1), 1-12.

Perin, C., Vuillemot, R., Fekete, J.D. (2013). SoccerStories: A kick-off for visual soccer analysis. *IEEE Transactions on Visualization and Computer Graphics*, 19(12), 2506-2515.

Rein, R., Memmert, D. (2016). Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. *Springerplus,* 5(1), 1410.

Sampaio, J., Macãs, V. (2012). Measuring tactical behaviour in football. *International Journal of Sports Medicine,* 33(5), 395-401.

Sarmento, H., Marcelino, R., Anguera, M.T., Campanico, J., Matos, N., Leitao, J.C. (2014). Match analysis in football: a systematic review. *Journal of Sports Sciences*, 32(20), 1831-1843.

Schuth, G., Carr, G., Barnes, C., Carling, C., Bradley, P.S. (2015). Positional interchanges influence the physical and technical match performance variables of elite soccer players. *Journal of Sports Sciences*, 34(6), 1-8.

Smith, P. (2018). How Belgium shocked Brazil 2-1 to reach World Cup semi-final. [online] [viewed 3/05/2019]. Available from: https://www.skysports.com/football/news/11096/11428909/how-belgium-shocked-brazil-2-1-to-reach-world-cup-semi-final

Stein, M., Janetzko, H., Breitkreutz, T., Seebacher, D., Schreck, T., Grossniklaus, M., Couzin, I.D., Keim, D.A. (2016). Director's cut: Analysis and annotation of soccer matches. *IEEE Computer Graphics and Applications*, 36(5), 50-60.

Stein, M., Janetzko, H., Lamprecht, A., Breitkreutz, T., Zimmermann, P., Goldlücke, B., Schreck, T., Andrienko, L., Grossniklaus, M., Keim, D.A. (2018). Bring it to the pitch: Combining video and movement data to enhance team sport analysis. *IEEE Transactions on Visualization and Computer Graphics*, 24(1), 13-22.

The Coaches Voice. (2018). Tactical analysis: Brazil 1 Belgium 2. [online] [viewed 23/04/2018]. Available from: https://www.coachesvoice.com/tactics-brazil-belgium-hazard-debruyne-fellaini-lukaku-neymar/

Tierney, P.J., Young, A., Clarke, N.D., Duncan, M.J. (2016). Match play demands of 11 versus 11 professional football using global positioning, system tracking: Variations across common playing formations. *Human Movement Science*, 49(1), 1-8.

Transfermarkt. (2018). World Cup 2018: Belgium-Japan: Lineups. [online] [viewed 23/04/2019]. Available from: http://www.espn.com/soccer/match?gameId=498149

Visser U., Drücker, C., Hübner, S., Schmidt, E., Weland, H.G. (2001). Recognizing formations in opponent teams. In P. Stone, T. Balch, G. Kraetzschmar (Eds.). RoboCup 2000: Robot Soccer World Cup IV. Lecture Notes in Computer Science. Berlin, Heidelberg, Germany: Springer.

Wallace, J.L., Norton, K.I. (2014). Evolution of the World Cup soccer final games 1966-2010: game structure, speed and play patterns. *Journal of Science and Medicine in Sport*, 17, 223-228.

Wilson, J. (2014). Inverting the pyramid: the history of football tactics. New York, NY, U.S.A.: Perseus Books Group.

Wu, Y., Xie, X., Wang, J., Deng, D., Liang, H., Zhang, H., Cheng, S., Chen, W. (2018). ForVizor: Visualizing spatio-temporal team formations in soccer. *IEEE Transactions on Visualization and Computer Graphics*, 25(1), 65-75.

Yiannakos, A., Aramats, V. (2004). Evaluation of the goal scoring patterns in European Championship in Portugal. *International Journal of Performance Analysis in Sport*, 6, 178-188.

Zauli, A. (2003). Soccer modern tactics: Italy's top coaches analyze game formations through 180 situations. Spring City, TN, U.S.A.: Reedswain Publishing

PART III – GENERAL DISCUSSION AND CONCLUSION

9.

GENERAL DISCUSSION

The goal of this thesis is to contribute to the domain of sports analytics by introducing new methods for dynamic and static sports analytics based on the Qualitative Trajectory Calculus. In Part I of this dissertation (Chapters 2, 3, 4 and 5), the proposed methods were introduced and explained in-depth. In Part II, three use cases were presented to evaluate the use of the proposed methods, two for the dynamic sports analytics method and one for the static sports analytics method. To conclude, Part III includes two chapters and aims at providing an overview of the main findings of the research. First, this chapter includes a general discussion of the research findings with respect to the research objective(s). It is divided in two subsections, where Section 9.1 presents an evaluation of the research objective(s) by means of a critical discussion of the main research outcomes for answering the posed research questions. Next, recommendations for further research are discussed in Section 9.2.

9.1 Discussion of the main research outcomes

This section presents a critical evaluation of the main research outcomes, based on the research objectives stated in Chapter 1 of this dissertation. Note that this discussion is limited to these research objectives and that the discussion of the results of the use cases for each of the specific domains is included in the respective chapters. Earlier, it was defined that the main research objective was:

The development of a methodology for (automated) sports pattern recognition through the analysis of complex movement interactions between different players based on the Qualitative Trajectory Calculus

As explained in Section 1.2.1, this main research objective encompasses two domains, *i.e.* dynamic sports analytics (team behaviour) and static sports analytics (team tactics). This division was detailed by discerning between two research questions (**R1** & **R2**). Before commencing the research, however, it was clear that if QTC would prove useful for analysing team behaviour and team tactics in sports, it most likely would have other applications in the domain sports analytics in a broader sense. For this reason, a third research question **R3** was posed. For clarity, the three research questions are mentioned once more:

R1 **Can the Qualitative Trajectory Calculus be used to (automatically) detect frequently occurring movement patterns between a certain number of players? (team behaviour)**

R2 **Can the Qualitative Trajectory Calculus be used to (automatically) detect tactical patterns in sports data? (team tactics)**

R3 **In which aspects of sports analytics could the QTC methodology be useful, besides analysing team behaviour and team tactics?**

As the three different research questions make use of the same method (QTC), the overall research could be defined as the finetuning and calibration of QTC for sports analytics. In this respect, following four more specific research questions were defined:

Ra **How to effectively compare different QTC-representations of sports intervals (team behaviour) or sports moments (team tactics) ? How to calculate similarities between different sports moments or intervals and decide whether two situations are similar? How can the results of the similarity calculation be validated?**

Rb **For which sports could QTC be used as a method for sports analytics?**

Rc **In what manner can QTC be enhanced to be more suitable as a method for sports analytics:**

- **What is the impact of the use of different QTC-variants?**
- **What is the impact of the use of static points?**
- **What is the impact of using different thresholds for the assignment of QTC-characters?**
- **What is the impact of the inclusion or exclusion of the positional data of the ball in the case of ball sports?**

Rd **Can QTC be used for more detailed movement analysis, *i.e.* analysis of individual athletes by description and comparison of parts of the body by means of QTC?**

Indeed, there is a partial overlap of the general with the more specific research questions, making it challenging to provide a discussion of their answers in an orderly manner. For this reason, the discussion is presented according to the general structure of this dissertation, where the method for dynamic sports analytics is first discussed (Chapter 3 of Part I and Chapters 6 and 7 of Part II) and then the method for static sports analytics (Chapter 4 of Part II and Chapter 8 of Part II). This discussion is thus structured according to the general research questions. The first two sections address **R1** and **R2**, respectively. As **R3** can be answered using both the dynamic as well as the static sports analytics method, it is included in those two sections. Similarly, the answers to the more specific research questions (**Ra**-**Rd**) are discussed in those two sections, as the finetuning depends heavily on the type of application (dynamic or static). The chronology of the more specific research questions is used to structure the next two subsections. Finally, there is a (short) subsection discussing the advantages and limitations of QTC for sports analytics.

9.1.1 **R1**: Can the Qualitative Trajectory Calculus be used to (automatically) detect frequently occurring movement patterns between a certain number of players (team behaviour)? – **Dynamic sports analytics**

The Qualitative Trajectory Calculus can indeed be used to analyse movement patterns between a certain number of players, as was demonstrated in Use Case 1. Furthermore, it can be used to analyse movement patterns of individuals where the MPOs are not individual players but rather points fixed on the human body (**R3**, **Rd**), as was demonstrated in Use Case 2. There are, however, many considerations that need to be taken into account before applying the proposed method for dynamic sports analytics.

Firstly, the proposed method has been successfully applied for movement pattern analysis in sports (football) and gait analysis (DCD analysis). These applications, however, can be seen as examples of pattern recognition (*i.e.* pattern matching) rather than pattern detection, as mentioned in the first research question (**R1**). The difference is that, as argued in Section 6.1, pattern detection can (linguistically) be seen as a data mining procedure (Jain *et al.*, 2000). Data mining in this case means that the dataset containing spatio-temporal data is analysed to find unknown movement patterns. In pattern recognition, however, a definition of the movement pattern of interest exists, reducing the complexity of the problem when compared to movement pattern detection procedures. Indeed, for the dynamic football use case, a predefined reference fragment was used to query the football dataset. Besides that, the football dataset was divided into target fragments of equal length that have temporal overlaps of fixed lengths. For the dynamic gait analysis, no predefined reference fragment was used, but the complexity of the problem was reduced by dividing the dataset into 72 fragments, each containing one step of a child. Consequently, distances between each of the steps were calculated for different marker combinations, aiming at finding the best combination of markers for discerning between two predefined groups of individuals (DCD or TD). Given the extra information used in both use cases, they can hardly be seen as examples of movement pattern detection, but rather as movement pattern recognition cases. Nevertheless, the use cases positively evaluate the use of the proposed method for discerning between similar and non-similar movement patterns. For this reason, it can be assumed that the proposed method can be used for movement pattern data mining purposes, on the condition that such analysis does not exceed the existing computational limitations of the current implementation of the Python program. Solutions to reduce the code complexity that could allow for more complex and exhaustive sports analyses were suggested in Appendix 2. Note that for movement pattern recognition with the proposed method, only the top-*k* of the results are interesting (see Use Case 1). This means that it is not meaningful to compare target fragments with each other that (both) have low similarities with the reference fragment, as the movement patterns that make up these target fragments can be vastly different, even if the similarity values of both target fragments with the reference fragment are (almost) identical. Similarly, for movement pattern detection, only fragments that have (very) high similarity values between them will contain movements with a high visual similarity.

Secondly, addressing the question posed in **Ra**, a novel method was introduced for the distance calculation between QTC-matrix sequences of different length (see Section 3.8). Its use has been demonstrated by a three-cushion example and by the results of the two dynamic use cases. The Levenshtein distance metric indeed allows for the calculation of movement patterns of different length and can be used to take differences of speed into account when comparing different fragments. However, limitations of the method were only addressed briefly in this dissertation. The main remaining question is to what extent the method can compensate for differences in length, and to what extent such a compensation is desirable. Indeed, sometimes differences in speed should be penalized, but to what extent? Many questions remain. An optimal outcome of the Levenshtein distance metric, for example, is not guaranteed when comparing relatively short fragments. An indel is unlikely to occur during the alignment of short QTC-matrix sequences because of its higher relative cost with respect to the total edit distance when compared to the alignment of longer fragments. Even when the compared fragments are all relatively long, their difference in length should not be too large if the Levenshtein distance metric is expected to produce a distance low enough to consider two such fragments to be similar. When this problem arises, the temporal resolution of the dataset can be downgraded to reduce the relative difference in length between two fragments. Surely, the outcome of the Levenshtein distance metric can be influenced by changing the weights and costs. There are, however, limits to its usability which remain unknown and are topic of future research (see Section 9.2).

Thirdly, addressing the question posed in **Rb**, there are no limitations of the applicability of the proposed method to different sports from a methodological point of view. At the start of this research it was assumed that only team sports could be analysed using the method, because of the necessity of at least two MPOs to make a description with QTC. By including static points, however, it becomes possible to analyse movement patterns of individual players, thereby broadening the applicability of the method to individual sports. Nevertheless, this dissertation is limited to the analysis in three-cushion billiards, football and walking/running. In fact, many more experiments were conducted on different sports in the light of this research. This includes experiments in, amongst other, basketball, golf, hockey, karate, samba dance and volleyball. From these, only the experiments that could be considered to meet the standards of a scientific paper, thus with a clear set-up, representative dataset and rigorous (successful) validation of the results, were included in this dissertation (see Section 1.2.2). The experiments that did not meet the above-mentioned standards were omitted from this dissertation, limiting the discussed sports to the above-mentioned three. This does not necessarily mean that the dynamic sports analytics method was not promising for that application or sport, but rather that, due to time restrictions, the experiment was not conducted to an extent that its results could be fully and correctly validated.

Regardless of the sports, a clear and valid (research) purpose is required, *i.e.* by trying to give an answer to the following question: “What are you trying to detect in the dataset and what is the desired outcome of the analysis?”. When the research purpose is ill-defined, results of the analysis cannot be expected to be useful from a sports perspective nor can they be validated positively. This occurred, for example, for an experiment during which movement patterns were analysed in volleyball. For the experiment, the (x,y,t)-coordinates of the volleyball players of one team were logged manually during one set using high resolution video images recorded with a temporal resolution of 25Hz. The experimental set-up made use of the strict sequential character of a volleyball match where one match consists of three to five sets. Each of those sets contains 25 successive rallies, which consist of a sequence of complexes. The number of complexes in one rally is not predefined, but a rally always ends with a complex where one of the both teams wins the point. A point can be won if the ball hits the floor (in or out the field) or when a fault is detected (Koch & Tilp, 2009). From the perspective of one team, a clear distinction exists between complexes of an attacking nature and complexes of a defensive nature. Given this important difference, the purpose of the experiment was to discern between the attacking and the defensive complexes based on the distances calculated between the QTC-descriptions of the different plays. Results of this early experiment (which was conducted before the experiments that are reported by means of use cases in this dissertation) were limited as it later turned out that attacking and defensive plays can each contain movement patterns with significant differences. The (x,y,t)-coordinates of the volleyball players, however, make up an interesting dataset that could in fact be analysed in a manner similar to the experimental set-up of Use Case 1. If the dataset would be analysed using such a pattern recognition approach where one complex is chosen as reference and the other complexes are defined as target fragments, the results would probably not be interesting from a sports perspective. The reason is that the volleyball players move with high degrees of freedom and that complexes can therefore contain vastly different movement patterns. As the dataset consists of only 58 complexes (attacking and defensive), chances of finding two highly similar complexes are very limited. This example clearly demonstrates the need for a clear and valid research purpose when conducting a dynamic sports analysis.

Fourthly, addressing the question posed in **Rc**, knowledge on the finetuning of the QTC method for dynamic sports analytics was gathered by conducting the above-mentioned experiments. In general, the optimal parameters for performing a dynamic sports analysis are heavily dependent on the specificities of the experiment/use case. For this reason, it is not straightforward to give a clear-cut overview of the parameters that need to be set for future experiments. Consequently, there is still a QTC domain expert required to conduct dynamic sports analyses at this stage of the research, even for the sports discussed in the use cases of this dissertation. There are, however, some general guidelines that one can bear in mind when performing movement pattern recognition/detection on a sports dataset using the proposed method. We describe these guidelines in the order of the sub questions of **Rc**. The optimal **QTC-variant** for an analysis, for instance, depends on the movements that are to be analysed. It can be advised to always start an analysis with the most basic QTC-variant (level one QTC_B), as this requires less computational power. Furthermore, the differences between the QTC-descriptions of different movements can be understood easier with the more basic variant, to assess whether QTC is useful for that application. When the relative speeds of the players/MPOs are of high importance, level two QTC_B is advised. If the results of the first analysis are not satisfactory or when the number of players/MPOs is relatively low (see Figure 3-7), it can be advised to switch to the more descriptive QTC_C-variant. Another reason for using the more detailed QTC-variant is when highly similar movements are compared. An example of this can be seen in the three-cushion billiards experiment described in Section 3.8.3, where QTC_B turned out to be insufficiently descriptive to discern between different types of opening shots, as all of the opening shots had very similar movements for the most part of their trajectories. For the experiment described in Use Case 1, on the other hand, QTC_B turned out to be sufficiently descriptive given the large number of degrees of freedom that football players can move on the field. Indeed, Feuerhake (2016) states that, because of the limited accuracy of tracking technologies and the vast number of possible positions (and thus movements) of players on the soccer field, it is almost impossible to find two fragments that contain identical movements. With respect to the limited accuracy of the data, it can be noted that QTC is a relative calculus, meaning that only the relative movements in two fragments need to be the same in order to be considered as identical. Adding to this qualitative nature of the calculus, both the temporal (as in Use Case 1) as well as the spatial resolution can be reduced to cope with the limited accuracy of the data. Following this line of thought for choosing the best QTC-variant, it could be assumed that the more detailed QTC_C would be used to describe the walking patterns of children in Use Case 2. Indeed, different steps on a treadmill could be assumed to be relatively similar for the different children. In the experiment, however, all combinations of markers were used to perform a series of analyses. It was assumed that the combination of 7 markers would put enough constraints on the movement characteristics of the QTC_B-description of the different steps. As the aim of the analysis was not to limit the number of markers to achieve the best results, the more basic QTC_B-variant was chosen. Nevertheless, a combination of three markers turned out to be the best to discern between the two groups of children. The analysis was performed again using QTC_C, but no lower classification errors could be achieved.

Continuing the discussion on the finetuning of the QTC method for dynamic sports analytics (**Rc**), **static points** have proven to be useful to help discerning between movement patterns occurring at different absolute locations, such as the three-cushion billiards example presented in Section 3.3. Furthermore, by including static points into the QTC-description of movement patterns, rotation and scale are taken into account. The optimal placement of the static points can be different for each experiment or analysis. Nevertheless, it can be advised to place the static points outside of the bounding box that surrounds all of the movements that are being compared. To give the absolute location a higher importance, more static points can be added to the QTC-descriptions of the movements. The downside is that the computational requirements increase as more static points are included. Note that if more than one static point is included, the QTC-relations describing the 'movement' between the static points should be neglected during the distance calculation between two fragments, as they consist of invariant 'o'-characters (see Section 3.3). This was done, for instance, for the experiment presented in Use Case 1, where four static points were placed on the corners of the football field. In Use Case 2, no static points were included because the coordinates of the markers were logged relatively to a fixed origin that was placed on the treadmill. Furthermore, the treadmill did not move in between the recordings of the different steps of the different children.

Different thresholds for the assignment of the 'o'-character were applied during the different experiments conducted for this dissertation, to see whether such an angular tolerance (α) could positively influence the capabilities of the dynamic sports analytics method. Results of analysing movement patterns with various different angular tolerances in sports were limited, given the issues mentioned in Section 3.5. For this reason, they were not included in any of the use cases reported in this dissertation. Consult Section 3.5 for a more detailed overview of the problems regarding the use of an angular tolerance for dynamic sports analytics. The impact of the **in- or exclusion of the ball** for (team) ball sports could not be studied thoroughly in this dissertation, as no dataset could be obtained that included accurate loggings of the movements of the ball. For all the experiments conducted for this research, the only dataset that included the (x,y,z,t)-coordinates of the ball was the football dataset used in Use Case 1. An assessment of the quality of this data, however, revealed that it was not possible to analyse the movements of the ball with the proposed method, as for many of the timestamps the dataset contained duplicated or empty cells for the balls' trajectory. Moreover, the use case lacked a clear research aim for which the inclusion of the ball in the QTC-description of the movement patterns would give an added value.

Fifthly, addressing the question posed in **Rd** as well as the more general question **R3**, QTC can indeed be used for a more detailed dynamic movement analysis of individuals by analysing the movements of parts of the human body. This was demonstrated by the experiment described in Use Case 2. While it can be argued that this use case is not a typical example of sports analytics, its results clearly demonstrate the capabilities of the dynamic sports analytics method to analyse the movements of the human body. Given the difficulties of obtaining gait data (as described in the first paragraph of Chapter 7), the use case investigated whether the proposed method is able to detect a gait pathology (DCD) for children walking on a treadmill. As walking and running are basic components of most sports, analysing such subtle gait differences can be seen as an application of sports analytics on an individual level. Applied to sports analytics in a stricter sense, the proposed method could, for example, be used for optimizing walking and running movements of athletes. In gymnastics, the proposed method could be used to compare several tries of forward or backward flic-flacs. This comparison could be used to compute a consistency index for the flic-flacs of a gymnast. Such an index could be used to optimize the gymnast's performance, as movement consistency was described to be essential in gymnastics (Grassi *et al.*, 2005). In golf, current methods for golf swing optimization are mostly quantitative and lack detection and feedback mechanisms of the most important differences between movements (Chun *et al.*, 2013).

When attaching markers on the body of a golf player and on the golf club, the proposed method could be used for the comparison of different golf swings with an 'optimal' golf swing, similar to what is done with state-of-the-art quantitative approaches (Ghasemzadeh *et al.*, 2009). Specific differences could be detected by counting deviant QTC-characters for a given set of markers. An experiment analysing golf swings with the dynamic sports analytics method was performed in the light of this PhD research (Versluys, 2018). Although the results were promising, validation possibilities were limited because of the lack of sufficient test persons and the limited number of golf swings. For this reason, the experiment is not included in this dissertation.

Besides the experiment on the human body movements presented in Use Case 2, an experiment on a Samba dance dataset (Chavoshi, 2014) shows promising results. Samba dance is a rhythmical dance that is characterized by the sequential repetition of similar body movements (Naveda & Leman, 2008). The dataset of the experiment contains the (x,y,t)-coordinates of 5 markers placed on the body of the samba dancers (one marker on the head and two markers on both hands and feet of the dancer). The three dancers (one teacher and two apprentices) each performed four repetitions of a samba dance piece that were logged with a temporal resolution of 25Hz. During these repetitions, the apprentices aimed at imitating the dance movements of the teacher as good as possible. For the analysis, each of the 12 repetitions are considered as fragments to compare. Using QTC_{B22}, it is possible to discern between the different dancers, as is shown in the hierarchical cluster in Figure 9-1. Furthermore, it can be seen that the teacher produced dance movements that were more constant over the different repetitions. Besides this, the dance movements of Apprentice A best resemble the movements of the teacher, but the different repetitions are characterized by a high variance. Apprentice B, on the other hand, produced relatively constant dance movements that, unfortunately, were very dissimilar from the dance movement of the teacher.

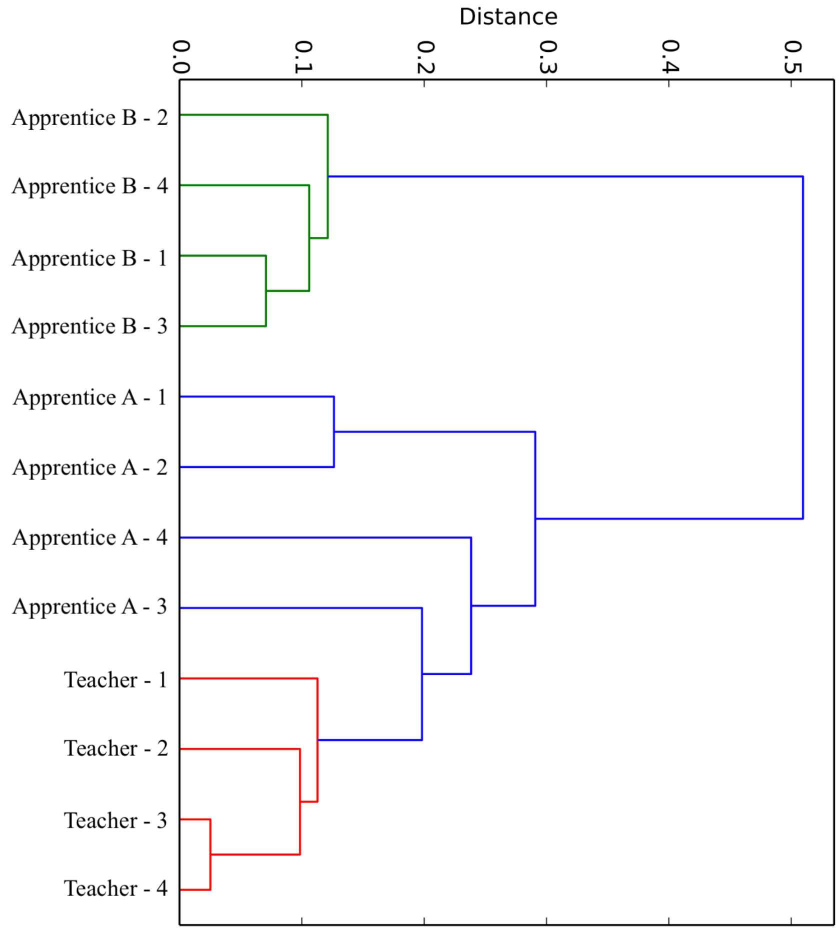


Fig. 9-1. Hierarchical clustering (UPGMA) of the different dance repetitions of one teacher and two apprentices using QTC_{B22}. Clusters with a distance lower than 1.25 are color-coded in red and green.

9.1.2 R2: Can the Qualitative Trajectory Calculus be used to (automatically) detect tactical patterns in sports data (team tactics)? – **Static sports analytics**

The Qualitative Trajectory Calculus can indeed be used to detect tactical patterns in sports data, as was demonstrated in Use Case 3. To this end, a new static variant, QTC_S, was introduced in Chapter 4 of this dissertation. Although considerably more efforts were allocated for answering **R1**, the proposed method for static sports analytics can indeed be seen as promising in this domain. There are, however, many considerations that need to be taken into account before applying the proposed method for static sports analytics, which are presented below.

Firstly, some remarks can be made regarding the formulation of the proposed method and the research question. It should be noted that although the method for static sports analytics compares configurations of MPOs at certain timestamps (from which it derives its name as “static”), it could be considered as a dynamic sports analytics method if every timestamp of a given dataset would be described with QTC_S and compared consecutively. However, as the method basically describes (and compares) point patterns at certain timestamps, it was named ‘static sports analytics’ in this dissertation. Moreover, a rather narrow vision of the applicability of the method was adopted when defining the research question. Indeed, the method can be used for analysing tactical patterns when applying it to team sports where the players (averaged) positions reflect the tactical intentions of the team (e.g. in football, basketball and hockey). The method can, however, also be used for the analysis of the position of different parts of the body during (individual or team) sports. This is a promising domain that was not considered while formulating the research questions and should be the topic of future research (see Section 9.2).

Secondly, addressing the question posed in **Ra**, no novel method for calculating the distance between two QTC_S-matrices was proposed in this research. Indeed, the distance calculation consists of a pairwise comparison of the QTC-relations that make up both QTC_S-matrices and is derived from the distance calculation method of the dynamic QTC-variants. It makes use of the conceptual distances between the QTC-characters, which is the only novelty with respect to distance calculation methods in previous research, and remains the only way of comparing two point configurations (see Section 4.2). The downsides are the necessity of the two point configurations to each contain two points, and the risk of over-averaging important positional differences due to the relatively large dimensions of the QTC_S-matrix.

Thirdly, addressing the question posed in **Rb**, the application of the proposed method for analysing different sports is mostly defined by the number of players. It is, for instance, not possible to describe the positions of individual MPOs with QTC_S without the use of static points (which are discussed in the next paragraph). Similarly, team sports with a low number of players (2-3) will need to be analysed by including static points to make sure that the QTC_S description is descriptive enough (see three-cushion billiards example at the end of Chapter 4). For team sports with a higher number of players, on the other hand, risks of over-averaging important individual positional characteristics when comparing different point configurations increase significantly, as the dimension of the QTC_S-matrix increases linear (quadratic) with every extra included player. Different weights could be assigned to the vectors in order to minimize this impact. On the downside, defining these weights is a time consuming effort that reduces the possible benefits of QTC_S as a method for automated static sports analytics. Nevertheless, it is possible to analyse the sports performance of individuals with the proposed method, by analysing the positions of markers on their body. As argued in Chapter 7, the data gathering is the limiting factor for performing such analysis at the moment. For this reason, current state-of-the-art sports applications are limited to walking and running in an indoor environment. More suggestions on the application of the static sport analytics method for analysing the human body are given in Section 9.2.

Fourthly, addressing the questions posed in **Rc**, it can be said that the amount of knowledge gathered on the finetuning of the QTC_S-method was limited because of the low number of experiments that could be conducted. In this dissertation, only the level one **QTC_S-variant** (having two characters in each cell of the QTC_S-matrix) was used. The reasons for this decision are the same as the ones mentioned in the previous paragraph, but the main aim was to reduce the amount of information stored in the QTC_S-matrices. The use of (a) **static point**(s) has been demonstrated by the three-cushion billiards example described in Section 4.3. As mentioned earlier, the number of added static points should be limited to the strict minimum. **Different thresholds** for the assignment of the QTC-characters were not used in the experiments, as some problems similar to the ones of the dynamic QTC-variants, are to be expected (see Sections 4.4 and 9.1.1). Since the datasets that were analysed with the static sports analytics method did not include tracked data of the ball, the impact of the **in- or exclusion of the ball** for team ball sports could not be studied thoroughly in this dissertation.

Fifthly, addressing the question posed in **Rd** as well as the more general question **R3**, it can be noted that the static sports analytics method can be used for a more detailed analysis of individuals by analysing the relative positions of parts of the human body at specific moments. Although this was not demonstrated by a use case in this dissertation, preliminary experiments demonstrate the usefulness of QTC_S for this purpose. Using the method, the spatial configuration of a series of markers on the body can be compared with the desired configuration at specific timestamps. This can be interesting for sports where gait and posture are of high importance and for sports that can be tracked (easily) in an indoor environment. In golf, for example, there are eight important moments during a golf swing for which ideal body postures can be described (Kelly *et al.*, 2010). The compliance of a golf player with the desired body postures at each of those moments consequently defines the quality of the shot resulting from that swing. An experiment, aimed at analysing these moments with the method for static sports analytics, shows promising results. The dataset of this experiment consists of the (x,y,z,t)-coordinates of two players who each perform five golf swings. The coordinates of fourteen markers on the body were logged with a temporal resolution of 250Hz, allowing to select the exact moments that were defined by Kelly *et al.* (2010). The (y,z)-positions of the fourteen markers at each of the 80 moments (10 shots) can be described and compared with level one QTC_S. The results are displayed in Figure 9-2 and show that, regardless of the player performing the swing, the proposed method can be used to discern between the eight different moments that occur during a golf swing. Furthermore, it can be seen that similar groups of moments, e.g. the moments of the set-up and the moments of the impact have small distances with respect to each other. Although both players have significantly different body shapes (e.g. height and weight), their respective moments were clustered together. The reason for this is the relative character of QTC_S. This is one of the main advantages of the static sports analytics method and enables a meaningful comparison of the positions of parts of the human body of multiple individuals. Unfortunately, the dataset does not contain enough test persons or golf swing repetitions to perform a more thorough analysis and validation. Furthermore, the lack of a QTC_S-description of the desired postures at each of the 8 moments impedes an assessment of the quality of the performed golf swings. This experiment, however, clearly demonstrates the capabilities of the static sports analytics method to perform analyses on the positions of parts of the human body.

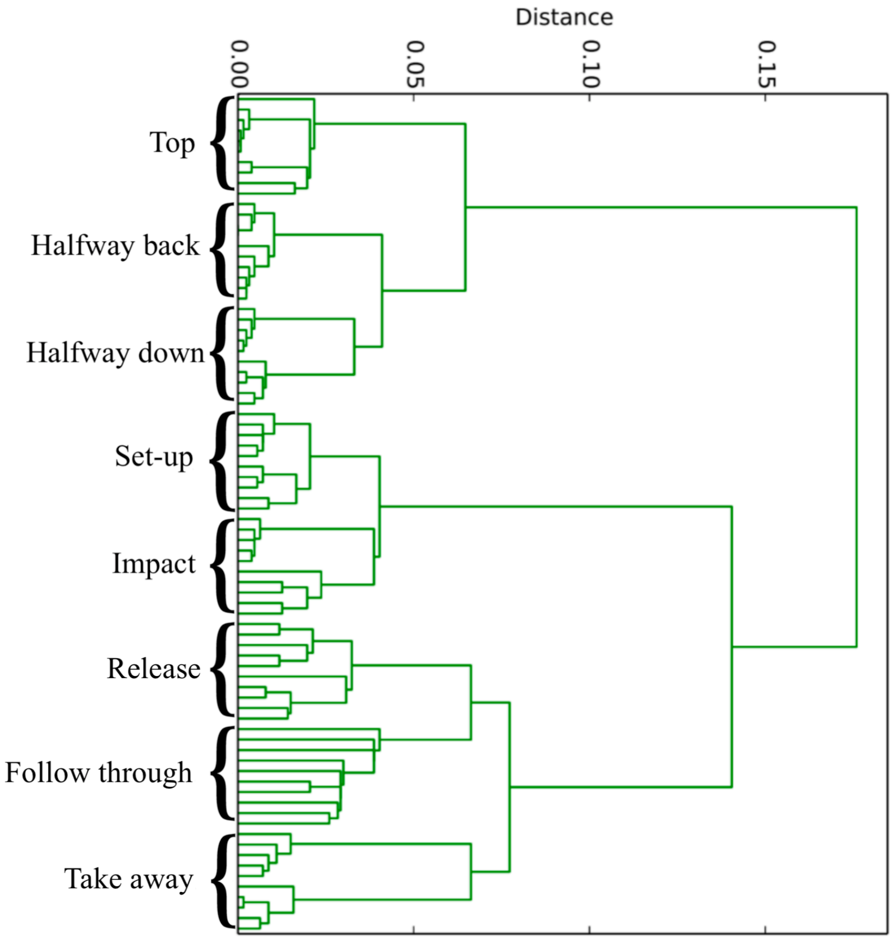


Fig. 9-2. Hierarchical clustering (UPGMA) of 80 different moments during 10 golf swings performed by two different golf players. Distances are calculated between the level one QTCS-representations of the positions of fourteen markers at each of the moments.

Finally, returning to the original phrasing of the research question **R2**, it is important to note the impact of the selection of the time intervals on the results of the tactical analyses when analysing team sports. Indeed, the players' positions are averaged for each of the time intervals before they are analysed with QTCS. In Use Case 3, for example, the time intervals of the official FIFA match reports are used to analyse the team tactics of the Belgian national football team (see Section 8.4). These match reports contain the positions of the players which are averaged for every period of 15, 45 and 90 minutes of the match. It is important to note that the results of such an analysis may differ significantly when different time intervals are used. Figure 9-3, for instance, shows the labels that are assigned to the positions of the 11 players of an anonymous team during the first half of a match using the time intervals of the official FIFA match reports. The match is selected at random from the first football dataset[11], and level one QTC_S is used for the analysis of the first half. As the dataset contains the (x,y,t)-coordinates of the players logged with a temporal resolution of 25Hz (see Section 8.2.1), it is possible to calculate the averaged positions for different time intervals. Figure 9-4, for example, shows the histogram of the labels assigned to the positions of the players that are averaged for every minute of the first half of the same match. Although Figures 9-3 and 9-4 both present a tactical analysis of the same data (the first half of the same match), their results differ significantly. Based on the results of Figure 9-3, it could be assumed that the team mostly played according to the 4-1-2-1-2(2) reference team formation. Looking at the more detailed results of Figure 9-4, however, the team most often played according to the 4-3-3(5) and the 4-1-2-1-2 reference team formations. Although in this general discussion section there is no space to discuss these differences in detail, this example clearly demonstrates the importance of the selection of the time intervals to conduct tactical analyses with the static sports analytics method.

ACTUAL TEAM FORMATION	LABEL	SIMILARITY WITH LABEL
0-15	4-1-2-1-2(2)	0.9299
15-30	4-1-2-1-2(2)	0.9349
30-45	4-3-1-2	0.8999
0-45	4-1-2-1-2(2)	0.9083

Fig. 9-3. The most similar reference team formations of an anonymous team for the positions averaged for each 15 and 45 minutes of the first half of an anonymized match.

[11] This is the dataset that was described in Section 8.2.1 and used both for Use Case 1 as well as for the experiments of Use Case 3.

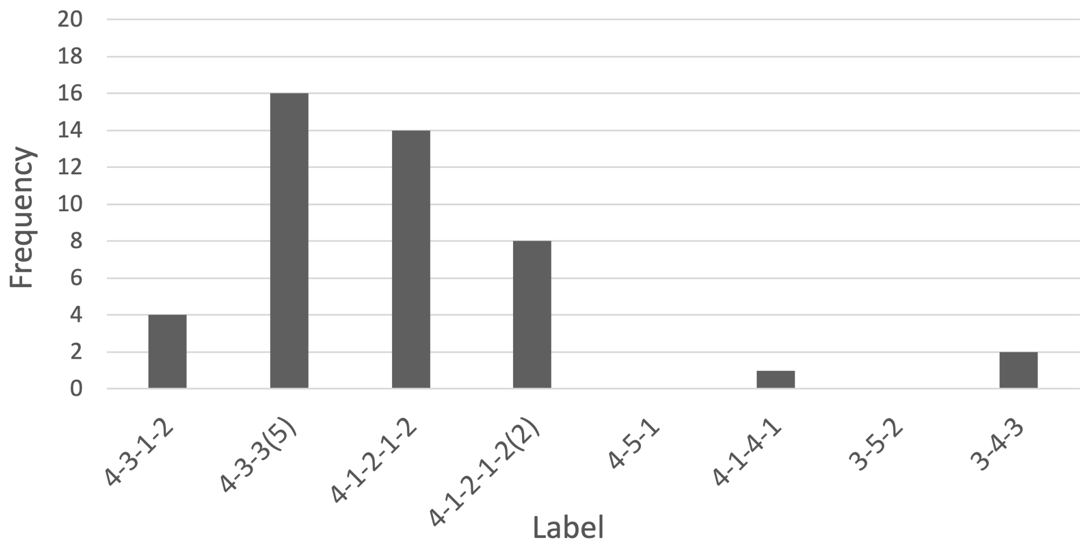


Fig. 9-4. Histogram of the labels assigned to the actual team formations of the first half of an anonymous match using a 1-minute average of the players positions.

9.1.3 Advantages and limitations of QTC

When compared with different qualitative spatio-temporal calculi and methods (see Section 2.1), QTC seems particularly suitable to analyse team behaviour and team tactics in sports. This was demonstrated by three use cases, but can also be discussed from a more theoretical viewpoint. For example, using RCC (Randell *et al.*, 1992) as a sports analytics method does not seem promising, as the studied objects are by definition non-overlapping MPOs. Because of this, all qualitative descriptions of sports movement patterns would consist of non-meaningful DC (disconnected) relations, making it impossible to discern between similar and non-similar movement patterns. Another qualitative calculus for which the use for sports analytics has been suggested, is REMO (Laube *et al.*, 2004). Using REMO, one could indeed describe and compare movement patterns in sports. There are, however, some important differences between REMO and QTC that need to be considered, which are demonstrated by a billiards example in Figure 9-5. On the right, the movement of four billiard balls during one time interval is shown. In the middle, the respective QTC_{B21} and REMO representations of this movement pattern are presented. The two qualitative representations correspond with a limitation of the possible movement directions for the four billiard balls during that time interval, which is shown on the right of Figure 9-5. It can be seen that REMO results in a limitation of the possible movement directions of the MPOs with a fixed angle of 45°. The strength of QTC, however, is that the movements of the MPOs are described with respect to each other, rather than compared to a fixed reference system. This means that, when multiple MPOs are described with QTC, the different constraints are combined. For the red and the white billiard balls in Figure 9-5, for example, this results in a possible movement direction with an angle < 45°. The movement direction of the black billiard ball, on the other hand, is described more strictly with REMO than with QTC. Because in sports analytics multiple time intervals are described, it can be assumed that each of the MPOs will be described very strictly with QTC for at least one time interval of a fragment. The combination of the constraints therefore results in a better description of the movement patterns in sports. Furthermore, the possible movement directions can be further reduced by including (more) static points in the QTC-description. Besides this, QTC can detect, if desired, rotated and/or translated movement patterns, which cannot be done using REMO.

Indeed, many other qualitative calculi and methods exist that can be compared with QTC. As the focus of this research was on the investigation of the use of QTC for sports analytics, relatively little efforts were made on the comparison of QTC with different calculi and methods. Instead, most efforts were aimed at conducting experiments with QTC in sports. As such, a more thorough comparison of QTC with other (qualitative) calculi that could be used for sports analytics remains a suggestion for future work (see next section).

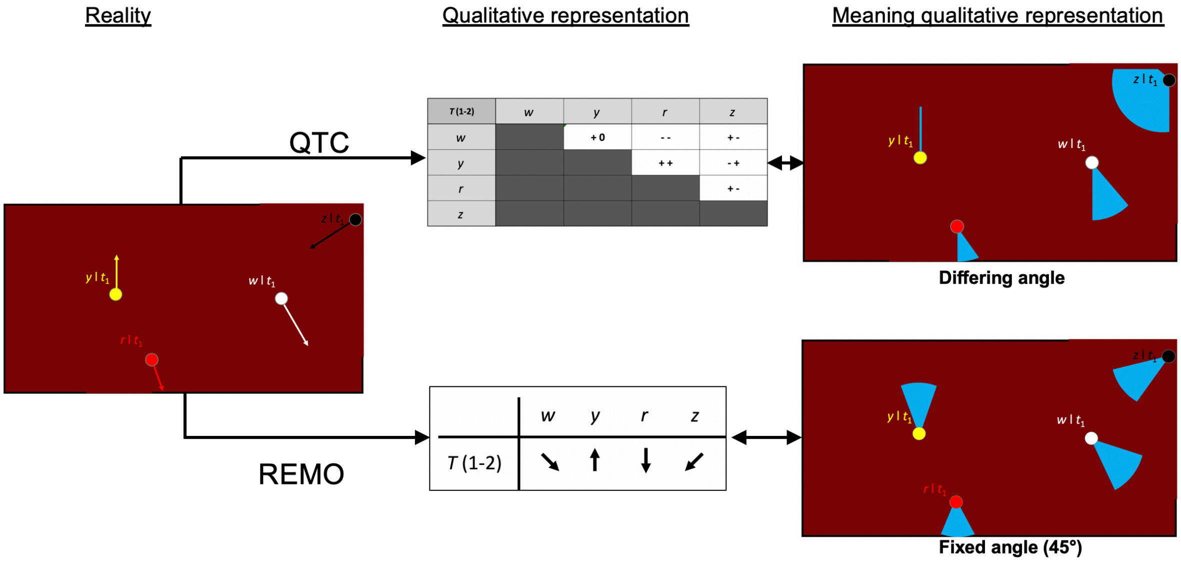


Fig. 9-5. The difference between QTC and REMO for analysing movement patterns in sports, illustrated by a billiards example including the movements of four balls during one time interval (left), the respective QTC and REMO representations of these movements (middle), and the meaning of the qualitative representations (right).

Obviously, there are limitations to the extent that QTC can be used as a method for sports analytics. It is most likely that QTC would be used as an add-on in existing applications for sports analytics rather than as a stand-alone solution. Indeed, QTC cannot not provide a full solution for the diverse questions posed in sports analytics. Sometimes this is because of the limitations of the method, while for a lot of sports analytics questions, an answer can be provided using simpler methods and techniques. In football, for example, one could try to analyse and detect the occurrence of off-side situations using QTC. A simple comparison of the (x,y,t)-coordinates of the players of the different teams, however, is sufficient to give a perfect answer to questions regarding off-side in football. The strength of QTC therefore lies in the subdomain of sports performance analytics, more precisely in the detection of reoccurring movement patterns and/or team formations in sports.

9.2 Recommendations for future research

In this section, possible avenues for future research are introduced. The aim of this non-exhaustive overview is to give suggestions for the most interesting future research efforts. First, recommendations for dynamic sports analytics are presented briefly, followed by some recommendations for static sports analytics. Note that section 9.1 also provides a series of possible avenues for future research.

9.2.1 Dynamic sports analytics

The method for dynamic sports analytics was discussed rigorously in Chapter 3 and illustrated by a three-cushion billiards experiment and two use cases. From a methodological perspective, however, a more thorough evaluation of the Levenshtein distance metric is required. The extent to which the method allows for the compensation of differences in length and/or pace of spatial movement patterns needs to be understood better in order to optimize the results of the dynamic sports analytics method. Furthermore, it could be investigated whether (and how) it is possible to describe and compare movement patterns with different (and differing) numbers of players with QTC. An important aim of future research should be a more thorough comparison of QTC with other (qualitative) calculi and methods for analysing movements patterns in sports (see Section 9.1.3).

The high complexity of the current implementation of the program results in high computational requirements for dynamical sports analyses and is undoubtedly the limiting factor of the research at this stage. Reducing the complexity of the program should therefore be one of the priorities of future research efforts, to support more complex analyses in the future. Suggestions on how to reduce the code complexity are provided in Appendix 2. Indeed, many additional applications of the method become feasible when more complex analyses could be supported by the program. Team behaviour in football, for instance, could be analysed by using more complex reference fragments that include more players. On top of that, multiple matches and player permutations could be taken into account. Furthermore, this would allow a data mining approach to find unknown movement patterns in larger football datasets and to gain more knowledge from the sports perspective. Similarly, such research could be conducted for team ball sports that are relatively similar to football (e.g. hockey, basketball, korfball). In basketball, for example, the sequential character of the match, including phases of defence and attack, could be used to define the different fragments of interest.

Future research on the movements of the human body could, for example, focus on measuring the flexibility of athletes. The flexibility of the human body can be defined as the ability to move body parts through a wide range of motion without causing non-reversible damage to the muscles or other tissue (Knudson *et al.*, 2000). Currently, flexibility of the human body is measured by a goniometer for every joint individually (Norkin & White, 2003). Unfortunately, this procedure is very labour intensive and time consuming, limiting its use to professional athletes only. Besides measuring movement angles of joints with a goniometer, there is currently no general test available that gives a representative value for the total body flexibility. Using movement data of markers placed on the body, however, the proposed method for dynamic sports analytics could be used to describe the flexibility of the human body. For this, the number of unique QTC-relations between (certain combinations of) the markers on the body could be counted, while asking the athlete to stretch his/her body in all possible directions (or as much as possible). To obtain the data needed for such an analysis, portable systems such as the Kinect (Microsoft Corporation, 2013) are increasingly used for human body movement research because of their simple and fast set-up (Li *et al.*, 2018). Moreover, using a combination of a colour RGB and an infrared camera, these portable systems are getting more precise in detecting movements of the human body. The Kinect, for example, allows for the projection of a skeleton with twenty body points in its field of view (Microsoft Corporation, 2013). Besides the suggestions provided in this section, there are numerous ways of applying the dynamic sports analytics method for experiments in different sports in the future.

9.2.2 Static sports analytics

The method for static sports analytics was discussed rigorously in Chapter 4 and illustrated by a three-cushion billiards example and a use case. From a methodological perspective, however, more research is required to study the impact of the different (proposed) methods for reducing the dimensions of the QTC_S-matrix. Furthermore, it could be investigated whether (and how) it would be possible to describe and compare configurations with different numbers of players/markers. An important aim of future research should be a more thorough comparison of QTC with other (qualitative) calculi and methods for analysing team formations in sports (see Section 9.1.3).

Future sports applications could aim at studying team tactics in football more thoroughly, for example by analysing team formations over longer periods and for multiple matches and/or seasons. Furthermore, results could be linked to sports performance data (e.g. match results) to create insights that could be used to enhance the team's performance. Similarly, other team sports with a suitable number of players and where the formation of the team is an important aspect of team tactics could be analysed. Examples are the analysis of hockey teams (Mitchell-Taverner, 2005), basketball teams (Lucey *et al.*, 2014), or parts of rugby teams (Hendricks *et al.*, 2013). On an individual level, there are various domains in sports analytics where the study of the spatial configurations of parts of the human body could be useful. Many possible applications exist, especially for sports where gait and body posture (at certain moments) are of high importance. When an ideal body posture can be described, as was discussed for a golf swing in Section 9.1.2, the method can be used to produce an assessment of the quality of the sports performance of an individual. Future research efforts could, for example, focus on the posture of the body at specific moments of a karate kick or in taekwondo (Gavagan & Sayers, 2017). Furthermore, the method could be used for the detection specific gait pathologies, as was done for the dynamic sports analytics for the detection of DCD in Use Case 2, by analysing the posture of individuals at specific moments of the gait.

REFERENCES

Chavoshi, S.H. (2014). *New directions in the analysis of movement patterns in space and time*. Ph.D. dissertation, Ghent University.

Chun, S., Kang, D., Choi, H.R., Park, A., Lee, K.K., Kim, J. (2013). A sensor-aided self coaching model for unlocking improvement in golf swing. *Multimedia Tools and Applications*, 72(1), 253-279.

Feuerhake, U. (2016). Recognition of repetitive movement patterns - the case of football analysis. *International Journal of Geo-Information*, 5(11), 1-20.

Gavagan, C.J., Sayers, M. (2017). A biomechanical analysis of the roundhouse kicking technique of expert practicioners: A comparison between the martial arts disciplines of muay thai, karate, and taekwondo. *PLOS One*, 12(8), 1-15.

Ghasemzadeh, H., Loseu, V., Jafari, R. (2009). Wearable coach for sport training: A quantitative model to evaluate wrist-rotation in golf. *Journal of Ambient Intelligence and Smart Environments*, 1(2), 173-184.

Grassi, G.P., Santini, T., Lovecchio, N., Turci, M., Ferrario, V.F., Sforza, C. (2005). Spatiotemporal consistency of trajectories in gymnastics: A three-dimensional analysis of flic-flac. *International Journal of Sports Medicine*, 26(2), 134-138.

Hendricks, S., Roode, B., Matthews, B., Lambert, M. (2013). Defensive strategies in rugby union. *Perceptual and Motor Skills*, 117(1), 65-87.

Jain, A.K., Duin, R.P.W., Mao, J. (2000). Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 22(1), 4-37.

Kelly, P., Healy, A., Moran, K.A., O'Connor, N. (2010). A virtual coaching environment for improving golf swing technique. In Proceedings of the 2010 ACM workshop on surreal media and virtual cloning (pp.51-56). Firenze, Italy: ACM.

Knudson, D.V., Magnusson, P., McHugh, M. (2000). Current issues in flexibility fitness. *President's Council of Physical Fitness and Sports Research Digest*, 3(10), 1-8.

Koch, C., Tilp, M. (2009). Analysis of beach volleyball action sequences of female top athletes. *Journal of Human Sport and Exercise*, 4(3), 272-283.

Laube, P., Van Kreveld, M., Imfeld, S. (2004). Finding REMO – Detecting relative motion patterns in geospatial lifelines. In P.D. Fisher (Ed.), Developments in spatial data handling, Proceedings of the 11th international symposium on spatial data handling (pp. 201-214). Berlin, Heidelberg, Germany: Springer.

Li, Q., Wang, Y., Sharf, A., Cao, Y., Tu, C., Chen, B., Yu, S. (2018). Classification of gait anomalies from kinect. *The Visual Computer*, 34(2), 229-241.

Lucey, P., Bialkowski, A., Carr, P., Yue, Y., Matthews, I. (2014). "How to get an open shot": Analyzing team movement in basketball using tracking data. *Disney research*.

Microsoft Corportation (2013). Kinect for Window. Human Interface Guidelines, pp.142.

Mitchell-Taverner, C. (2005). Field hockey – techniques & tactics. Champaign, IL, U.S.A.: Human Kinetics.

Naveda, L., Leman, M. (2008). Representation of Samba dance gestures, using a multi-modal analysis approach. In Proceedings of the 5th enactive international conference (pp. 68-74). Pisa, Italy.

Norkin, C.C., White, D.J., (2003). Measurement of joint motion: A guide to goniometry. Philadelphia, PA, U.S.A: Davis Company.

Randell, D., Cui, Z., Cohn, A.G. (1992). A Spatial Logic Based on Regions and Connection. In B. Nebel, W. Swartout, C. Rich (Eds.), Proceedings of the 3rd International Conference on Knowledge Representation and Reasoning (pp. 165-176). San Mateo, CA, U.S.A: Morgan Kaufman.

Versluys, C. (2018). *De Qualitative trajectory calculus als methode voor de analyse van de golfswing*. Master thesis dissertation, Ghent University.

10.

GENERAL CONCLUSION

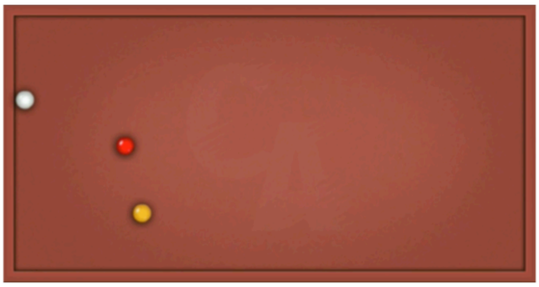
Interest in sports analytics has seen a significant growth in the last decade. In the subdomain of sports performance analytics, various methods were introduced, ranging from psychological and sociological assessments to more individual statistics that describe an athletes' respiration rate, heart beat or even nutrient intake ratio. Despite the vast variety of sports and their characteristics, most (team) sports consist of an important mutual component: spatial movement through running or walking. The current research is situated in the domain of sports analytics and aims at developing a methodology for (automated) sports pattern recognition, by analysing the complex spatial movement interactions between players. The proposed methodology is based on the Qualitative Trajectory Calculus (QTC), a geographical calculus describing the relative movement between Moving Point Objects (MPOs). As movement pattern recognition in sports can be applied both on movement data with a fine temporal resolution as on movement with a coarse temporal resolution (by analysing averaged positions), the distinction was made between a dynamic sports analytics method and a static sports analytics method. The first can be used to analyse spatial movement patterns of one or multiple players (team behaviour) and makes use of traditional QTC-variants. The latter can be used to analyse the spatial configuration of multiple players at specific moments in time or averaged over time (team tactics) and makes use of a novel QTC-variant named QTC_S. Both methods were described rigorously in this dissertation, in two separate chapters. Theoretically, both methods can be enhanced by a series of adaptations, including the use of static points, different QTC-variants, different thresholds for the assignment of QTC-relations or by the in- or exclusion (player) permutations in the similarity analysis. To illustrate the use of both methods for sports analytics and their suggested adaptations, they were exemplified by a series of three-cushion billiards examples. Three cushion billiards is a variant of the more popular carom billiards and was chosen as an example because of its simplicity. In three-cushion billiards, the number of objects are limited and the movements of the balls are only defined by an initial shot and the laws of nature. The absence of psychological or human factors influencing the trajectories of the balls after the shot has been taken reduces the complexity and allows to evaluate the use of the proposed methods for dynamic and static sports analytics before switching to more complex sports situations.

Besides these basic examples, a series of sports experiments were conducted to investigate the use of the proposed methods for specific sports and to evaluate the proposed adaptations in this respect. Besides the original aim of applying the methods on team sports, their use for analysing more detailed movements of individual athletes was investigated by a series of experiments. Experiments were conducted in different sports including basketball, football, golf, hockey, samba, tango, volleyball, walking and running. Based on the three-cushion billiards examples and the experiments, some conclusions can be drawn. For the dynamic sports analytics method, it can be concluded that static points are useful for taking the absolute location, the rotation and the scale of movement patterns into account. Similarly, the use of different QTC-variants can be evaluated positively for the dynamic sports analytics method, although the basic variant QTC_B can be preferred over QTC_C. Only when the basic variant does not deliver the expected results, more descriptive QTC-variants should be used. The use of different thresholds for the assignment of the QTC-relations, on the other hand, is not advised for the dynamic sports analytics method because of methodological reasons. Finally, permutations could be taken into account when comparing different movement patterns, although their use was limited in this dissertation because of their high demand of processing power. For static sports analytics, the use of the proposed adaptations is limited because of the large dimensions of the QTC_S-matrix, even when describing the spatial configuration of a limited number of players or MPOs. Based on the experiments, the use of static points can be evaluated positively for the static sports analytics method, but only when the number of additional static points is limited. Similarly, no different QTC_S-variants were described in this dissertation. The application of different thresholds for the assignment of QTC-relations is not advised for the static sports analytics method. Indeed, because the QTC-relations of the QTC_S-matrix are calculated similar to the relations of QTC_B, similar conclusions can be made as for the dynamic sports analytics method. Using (limited numbers of) permutations is very useful when comparing player configurations in team sports using the static sports analytics method, but less for analysing the movement of the human body. In this dissertation, however, the number of permutations and players was limited for the static sports analytics method because of the large dimensions of the QTC_S matrix and the high computational requirements. In general, it can be concluded that at this stage of the research, it remains necessary to consult a QTC-domain expert before conducting sports analyses using the proposed methods, even for the sports studied in the described experiments. Besides this, an optimization of the current Python implementations is necessary before performing more complex analyses in the future.

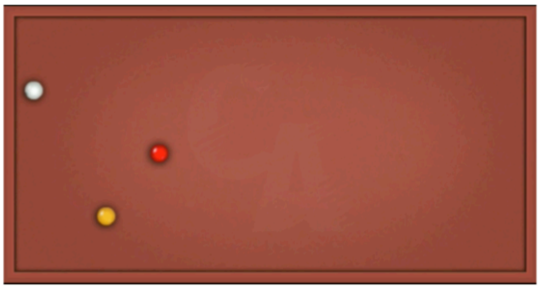
The experiments that included a clear set-up, a representative dataset and a rigorous (successful) validation of the results were included as use cases in the dissertation. Based on the results of these use cases, conclusions can be drawn for the use of the proposed methods for specific sports. The first use case presents an application of the dynamic sports analytics method for the recognition of basic spatial movement patterns in football. It clearly shows that the proposed method allows a user to select a movement pattern as a reference fragment and to search for similar movement patterns in one or multiple matches. It is important to mention that, especially in the case of football, only the fragments that have very high similarities with the reference fragment should be considered as meaningful results. Furthermore, the method could be applied easily the for movement pattern recognition in other team sports that have similar characteristics such as basketball or hockey. The second use case demonstrated that the proposed method for dynamic sports analytics can be used to analyse the more detailed movements of parts of the human body. More specific, it allows for the detection of gait pathologies, such as DCD , by describing and comparing the walking patterns of multiple individuals. For human body movement analysis, best results are expected when a distinctive reference movement or clear distinction between different groups of movements is available (e.g. a perfect golf swing or karate kick). The third case showed that the static sports analytics method can be used to describe and analyse the team formations of a whole team or groups of players in football. This shows that the method can be used to assess team tactics in football and should be easily applicable on other team sports with similar number of players where team formations reflect the team tactics.

Appendix 1
End positions of the three-cushion billiards opening shots

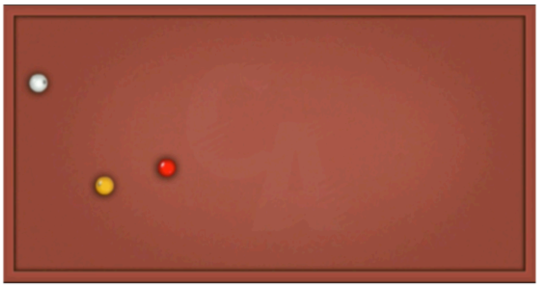
1 R



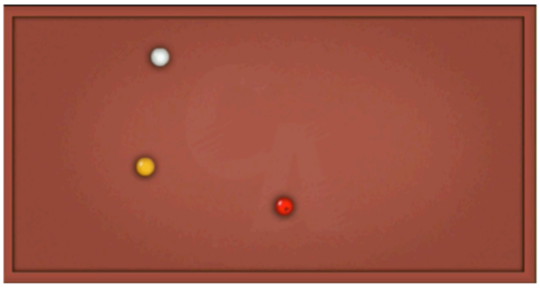
2 R



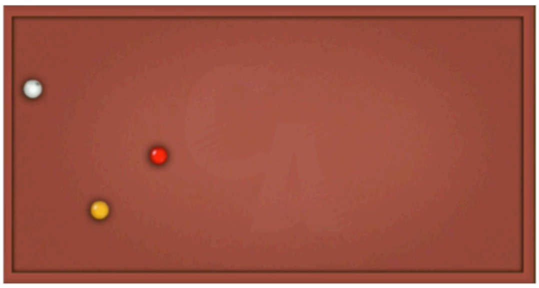
3 R



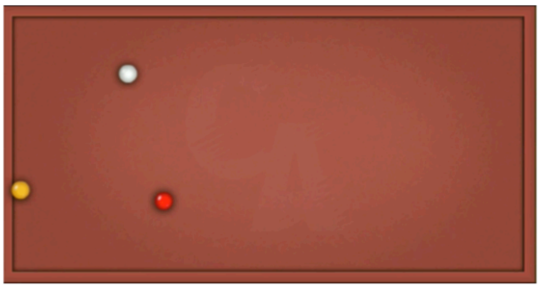
4 L



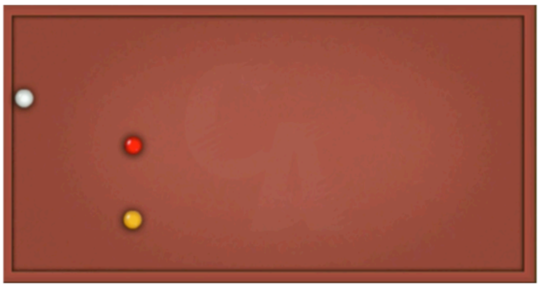
5 R



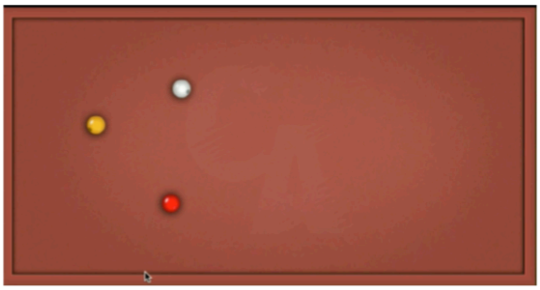
6 S



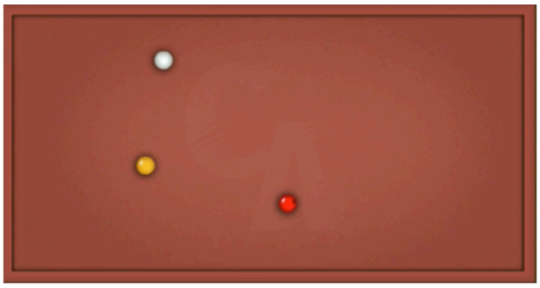
7 R



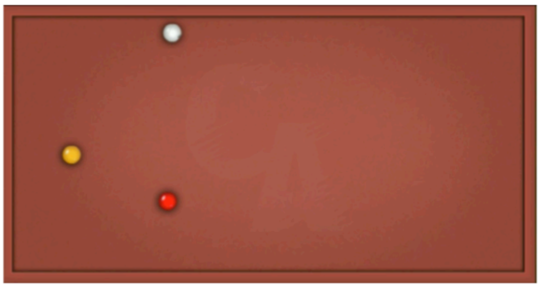
8 S



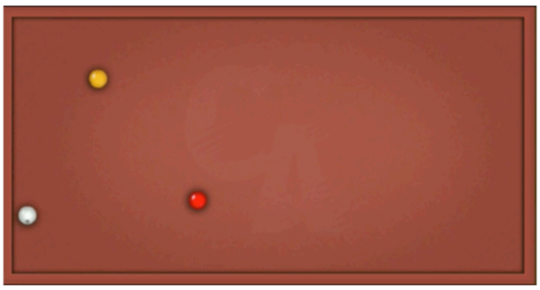
9 L



10 S

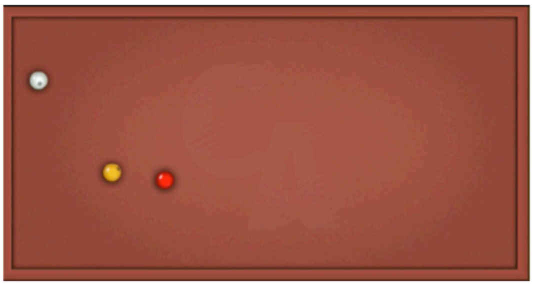


11 S

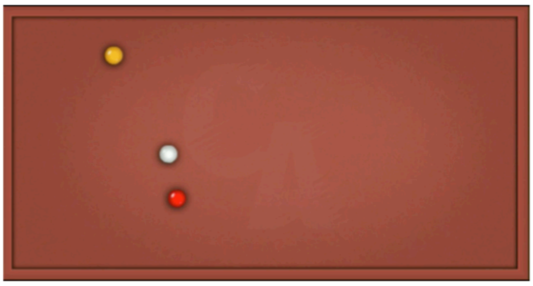


12 R

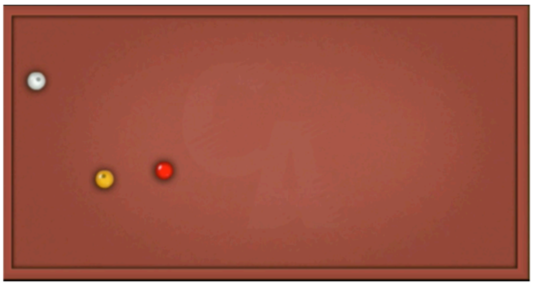
13 R



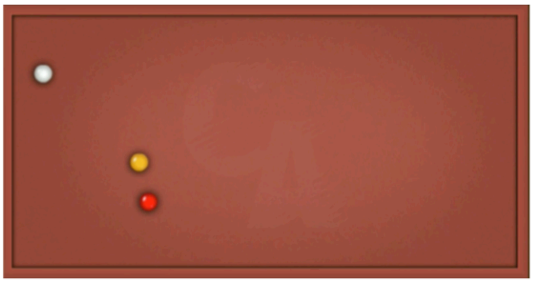
14 S



15 R

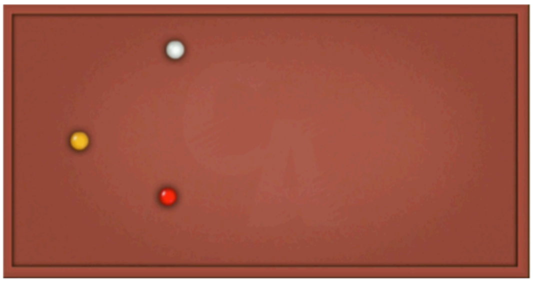


16 R

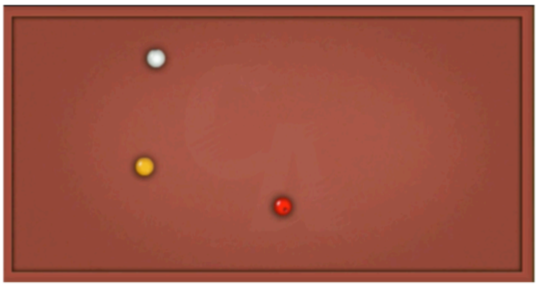


17 S

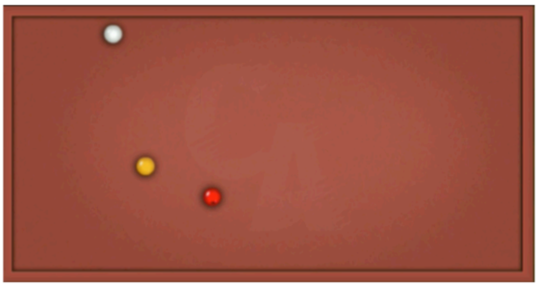
18 S



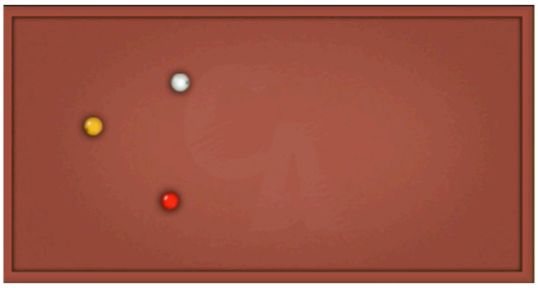
19 L



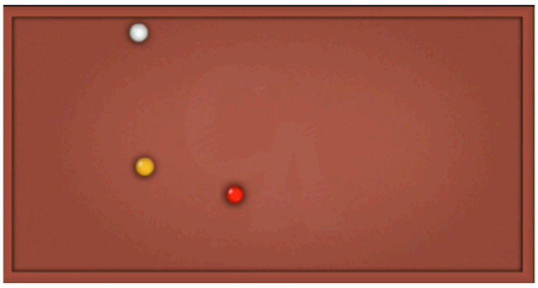
20 L



21 S

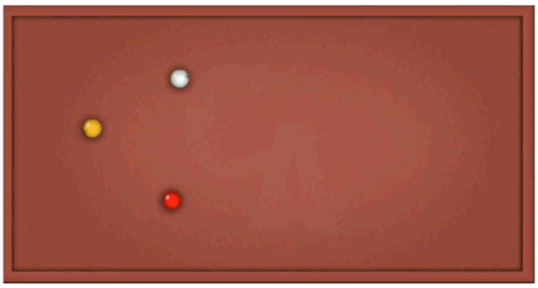


22 L

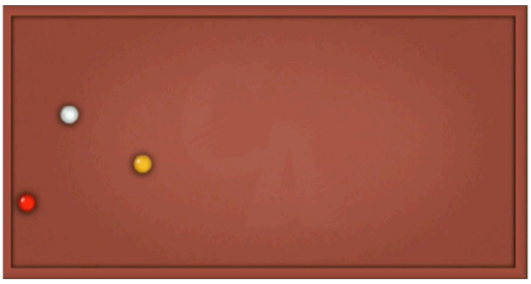


23 R

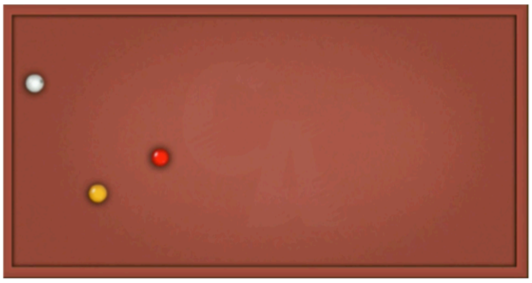
24 S



25 R



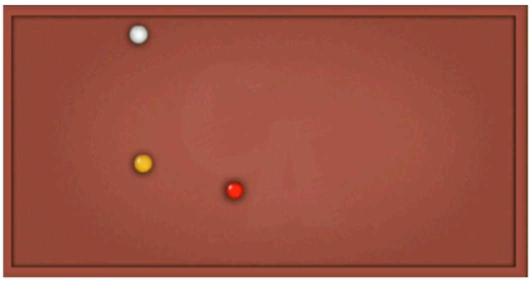
26 R



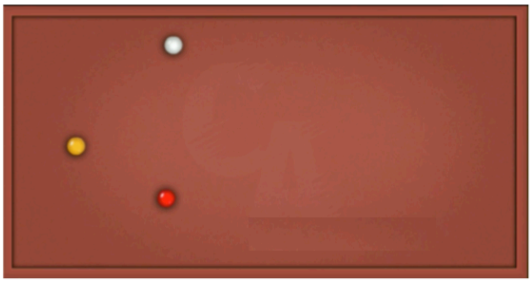
27 R

28 R

29 L

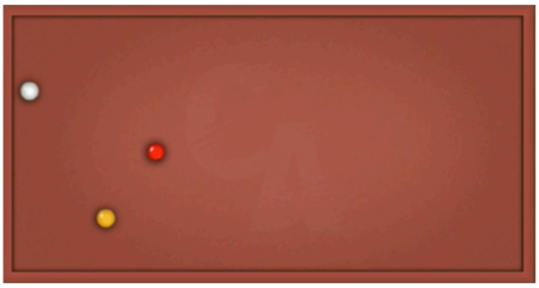


30 S

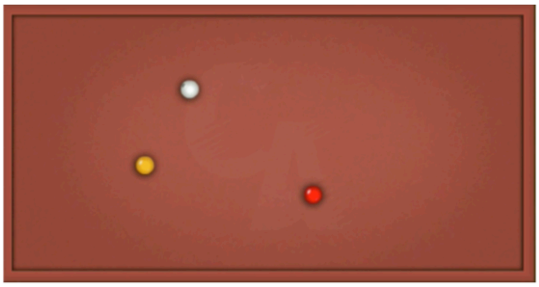


31 R

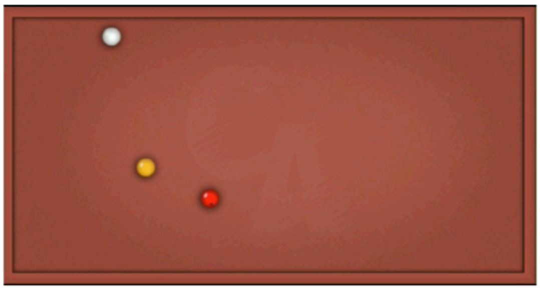
32 R



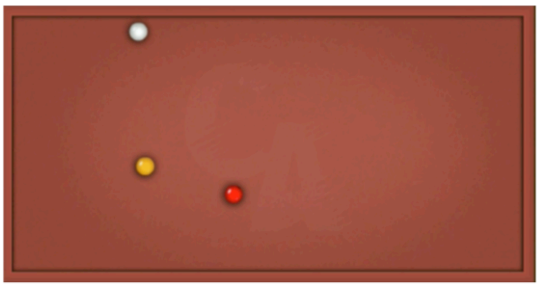
33 L



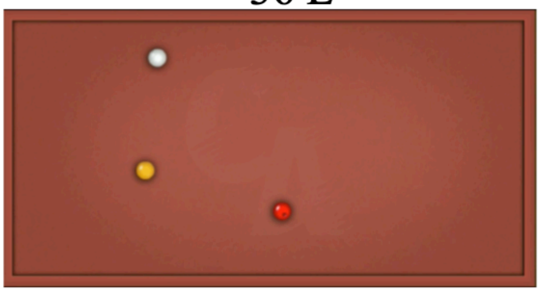
34 L



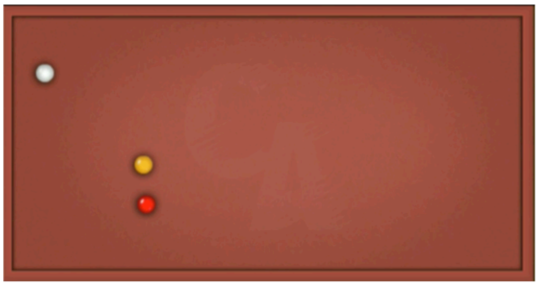
35 L



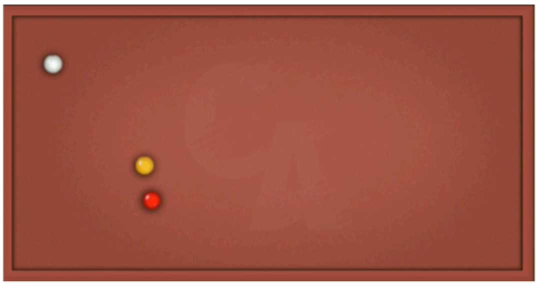
36 L



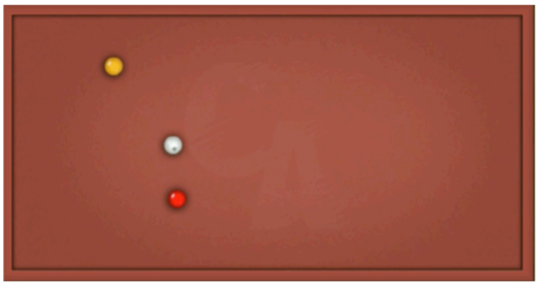
37 R



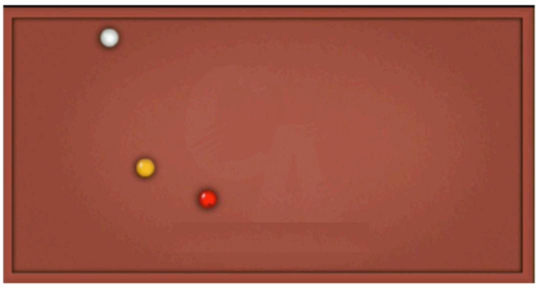
38 R



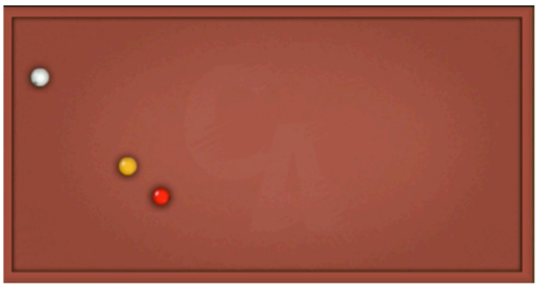
39 S



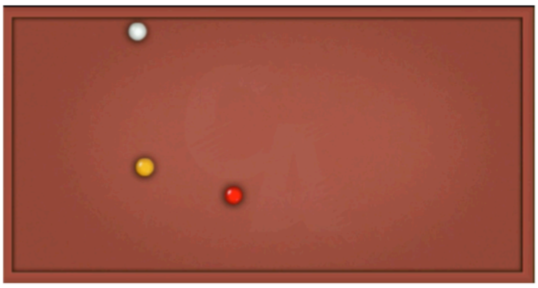
40 L



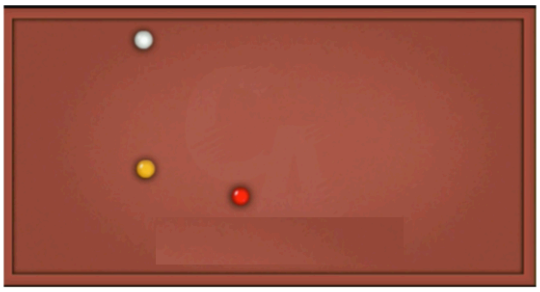
41 R



42 L

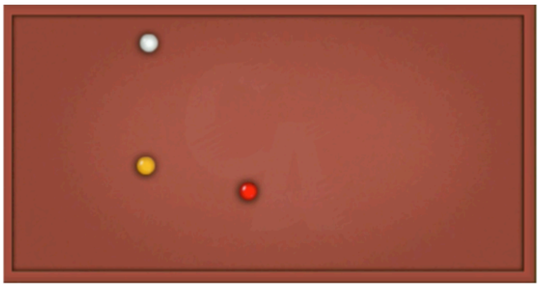


43 L

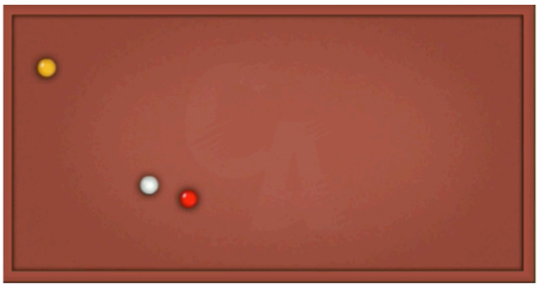


44 R

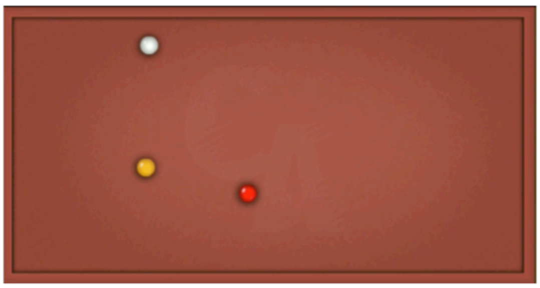
45 L



46 S

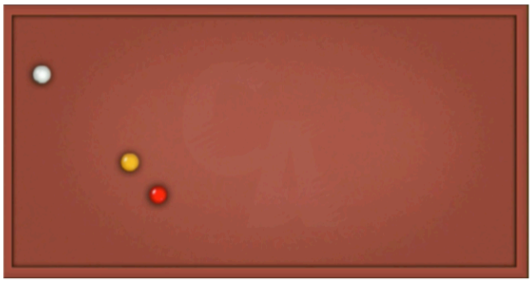


47 L

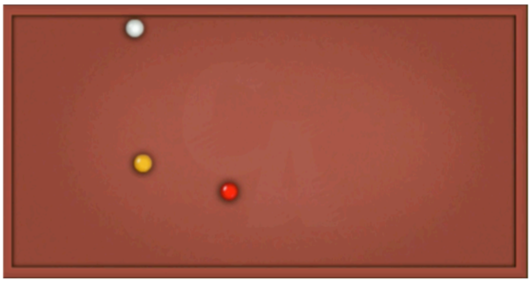


48 L

49 R

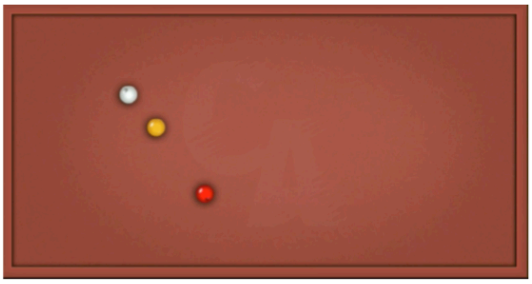


50 L

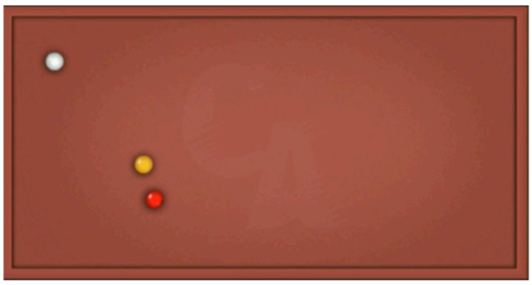


51 R

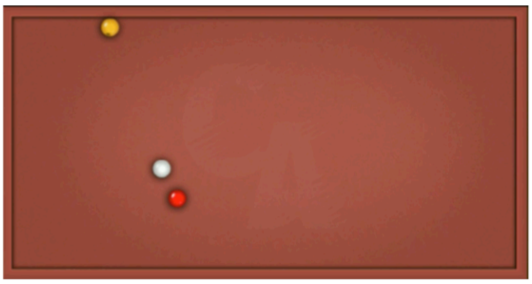
52 S



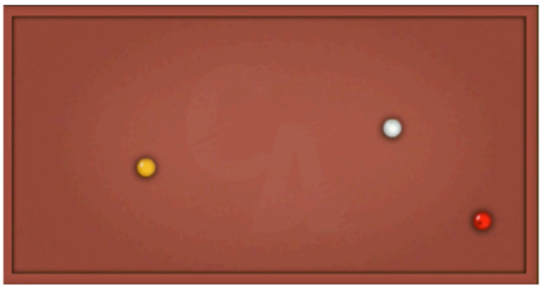
53 R



54 S



55 L



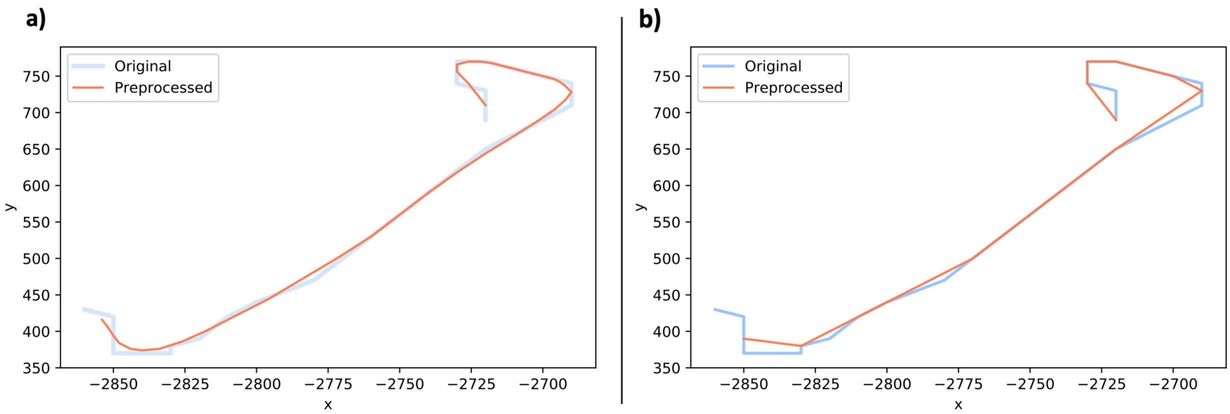
Appendix 2
Optimization of the Python programs

The Python programs make use of exhaustive, brute force techniques. This means that there is no optimization of the performed calculations, *i.e.* all of the QTC-relations are calculated and compared consecutively. This can result in enormous calculation times, especially for dynamic sports analytics[12]. There are, however, various techniques that could be used to reduce the programs' complexity and optimize their performance thereby reducing the computing power requirements for a given analysis. Besides enabling faster calculations, these techniques could support more complex calculations on bigger datasets. In the light of this dissertation, a master thesis topic was designed that stem from the necessity of this optimization. This master thesis was supervised by the author of this dissertation. Standaert (2018) investigated which techniques could be used for the optimization of the dynamic QTC sports analytics program. Given the importance of this research, a review of his findings are presented concisely in this section, alongside additional thoughts. It is important to note that this section presents a theoretical exploration of the optimization techniques solely for level one QTC_B. The application of the introduced techniques were not yet investigated for other QTC-variants, neither were they implemented in the Python programs. Standaert makes the distinction between techniques that can be applied at the preprocessing stage and techniques that can be applied during the actual QTC-analysis. They are presented separately in the following two sections. To conclude, a third subsection includes some additional (preliminary) ideas on the optimization of both the static as well as the dynamic sports analytics method.

[12] Dynamic QTC sports analytics is typically used for comparing numerous, often overlapping, time intervals that are described with QTC. Since between each two timestamps of each such time interval a QTC-matrix is calculated, large numbers of QTC-relations and matrices need to be computed and compared. Static QTC sports analytics, however, is typically used to compare a more limited number of timestamps, where one timestamp accounts for one QTC matrix and overlap is by definition impossible.

Preprocessing

During preprocessing, a series of steps can be applied for optimization before conducting the actual dynamic QTC-analysis. The first preprocessing step can be performed on the coordinate domain, by applying a moving average on the (*x*,*y*,*t*)-coordinates of the MPOs of the fragments to compare. This step is especially useful for analysing datasets that are prone to outliers. Subsequently, subsampling can be applied on the (*x*,*y*,*t*)-coordinates of the fragments, to reduce the temporal resolution of the dataset. The two steps of the coordinate domain preprocessing are shown in in the figure below, in which a moving average of 10 timestamps (a) and a subsampling of 25Hz to 10Hz (b) are performed on a fragment of a football dataset (see Chapter 6).



Coordinate domain preprocessing by applying a moving average of 10 timestamps (a) and a subsampling to reduce the temporal resolution of 25Hz to 10Hz (b) on a football fragment. (after Standaert, 2018)

The second step of preprocessing can be performed in the QTC-domain, by transforming the QTC-matrix-sequences representation of a fragment into compressed string representations. To do so, the QTC-representations of each fragment should be noted as sequences of tuples for each pair of MPOs, rather than the more conventional QTC-matrix-sequences notation (see Section 3.1). Consequently, each tuple can be converted into a letter representation by applying the mappings shown in the Table below. Each tuple sequence can thereafter be written as a string by putting the letters in sequence. As a basic example, a fragment of two MPOs *k* and *l* whose locations are logged for 11 timestamps can be presented (a). Normally, this fragment is described by means of a QTC_B-matrix sequence for level one QTC_B (b). Reducing the complexity of the notation, it can be noted by following sequence of tuples "(0, +), (0, +), (0, +), (0, +), (0, -), (+, +), (+, +), (+, +)" shown in c. This sequence of tuples can thereafter be converted into the string "KKKKULLL" (Figure 5-6d). The big advantage of using this string notation is that local repetitions of letters can be compressed using compression techniques.

Mapping table for the QTC-domain of level one QTC_B. (after Standaert, 2018)

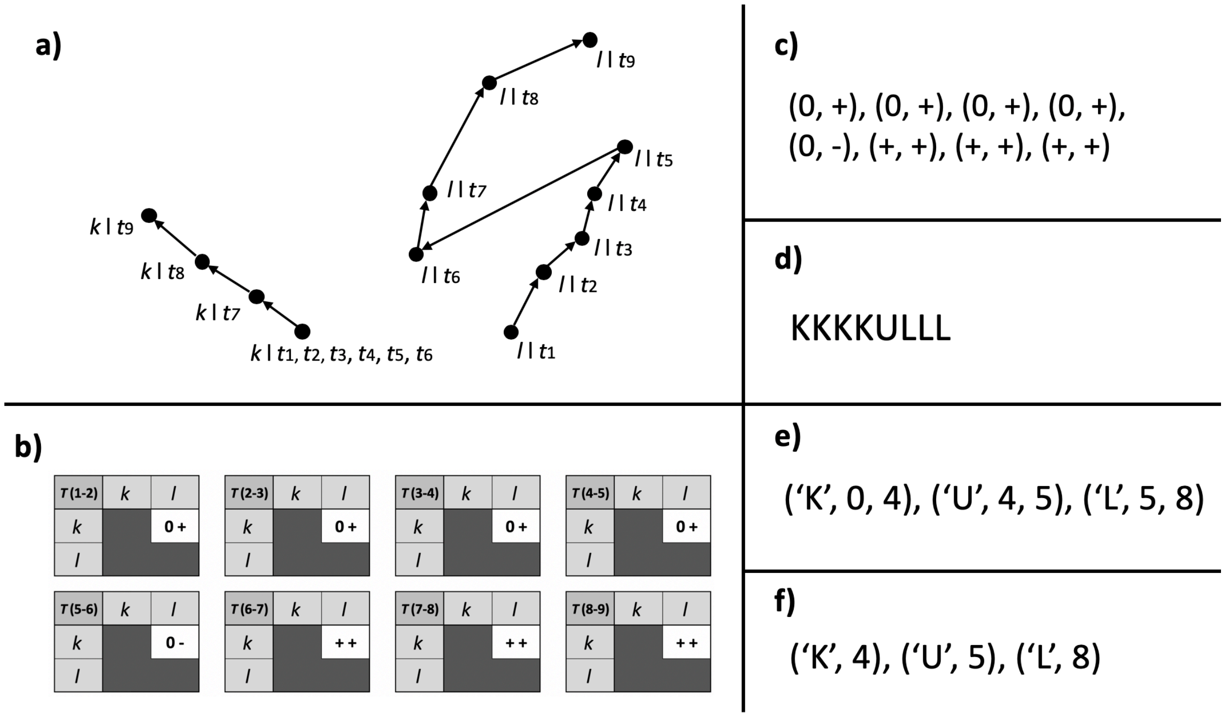
First character	Second character	Mapping
0	0	A
0	+	K
0	-	U
+	0	B
+	+	L
+	-	V
-	0	C
-	+	M
-	-	W

The first compression technique that can be applied is Low Level Compression (LLC), which is an example of a run-length encoding compression technique (Robinson & Cherry, 1967). LLC removes repetitions of letters by retaining only the start and end timestamp for each character of the string (Standaert, 2018). Following this procedure, the string notation of the example fragment "KKKKULLL" can be compressed into '('K', 0, 4), ('U', 4, 5), ('L', 5, 8) (d). LLC is a form of lossless compression since no information is being lost. Using LLC, Standaert (2018) reports compression ratios[13] ranging from 41% to 4% for two samples of datasets[14] that are used in the use cases in the second part of this dissertation. The differences between the compression ratios of the different datasets can be explained by different characteristics of the movements (e.g. different

[13] Compression ratios are expressed by the new size (compressed string notation) of the QTC-description divided by the old size (original string representation). Hence lower the compression ratios indicate more compression of the QTC-descriptions.

[14] 41% for the sample of the football dataset (see Section 6.3.1) and 4% for the sample of the gait treadmill dataset (see Section 7.2.1).

levels of movement variety or speed) of the different datasets. When a dataset contains more repetitive movements (and thus more repetitive QTC-relations), compression ratios will be lower. Going beyond his approach, it is possible to further reduce the size of the QTC-descriptions of the fragments, as the inclusion of both the start as well as the end timestamp of each letter is not necessary. Only including one of both timestamps is sufficient to ensure a lossless compression of the QTC-description, enabling even further compression and lower compression ratios. This is shown for the example fragment in (f). Alternatively, Standaert (2018) did propose another technique for further compression of the QTC-descriptions by applying High Level Compression (HLC) on the LLC-compressed fragments. The HLC technique, however, is a form of lossy (or irreversible) compression, that is not suitable for dynamic QTC sports analytics, at this stage of its development.



QTC-domain preprocessing on a fragment containing two MPOs k and l whose locations are logged at nine consecutive timestamps (a), the level one QTC_B-matrix sequence notation (b), the sequence of tuples notation(c), the string notation (d), LLC applied on the string notation (e) and the LLC using only the end timestamp (d).

While the coordinate domain preprocessing is optional, the QTC-domain preprocessing is necessary to apply the proposed optimized fragment comparison, which will be presented in the next section. A general remark is that the application of a moving average before LLC will greatly favour the compression and will result in significantly lower compression ratios (thus more compression). It is necessary to take this into account for a correct interpretation of the compression ratios.

Optimized fragment comparison

In the previous section, techniques were introduced to reduce the size of the sports data and the QTC-descriptions of the different fragments. A different approach for optimizing the dynamic QTC sports analytics program, however, is optimizing the way in which the different functions of the program (see the flowchart shown in Figure 5-1) are called upon and executed.

Except the compression of datasets (coordinate domain compression) and the compression of the QTC-descriptions of the fragments (QTC-domain compression), there is little opportunity for optimizing the first ten steps of the flowchart (Figure 5-1). More promising, however, are the abilities of optimizing the brute force 'Calculate_distance_between_fragments()'-function. In this function, distances between the QTC-matrix sequence representations of all the fragments are calculated, which is mostly done using the Levenshtein distance metric. For optimizing this procedure, at first the edit distance between the LLC compressed notations of both fragments can be calculated. Furthermore, an early stopping mechanism could be used to terminate the calculation of the edit distance between two LLC compressed fragments as soon as the intermediate distance exceeds a predefined threshold γ_1. In this way, the distance calculation is terminated before the total edit distance is reached (Standaert, 2018). Consequently, if the edit distance between the two LLC compressed fragments is calculated (meaning the edit distance is below the predefined threshold γ_1), the edit distance between the original QTC-matrix sequences of the fragments can be calculated. Yet again, the calculation of this distance can be stopped by adding a stopping mechanism that terminates the calculation of the edit distance once the intermediate distance exceeds a predefined threshold γ_2. Since, in a movement pattern recognition or detection setup, ultimately a user is only interested in highly similar fragments, such stopping mechanisms will eliminate unnecessary calculations of cells in the edit distance matrix (see Figure 3-12) and will optimize the dynamic QTC sports analytics programs' performance. Standaert (2018) goes even further, by adding a preliminary step where the edit distance calculation between two fragments is performed on their HLC compressed notations and the procedure mentioned in the previous paragraph is only called upon if this edit distance does not exceed a third predefined threshold γ_3. To give an idea about the optimization rates, Standaert (2018) reported a reduction of the (average) calculation time of a dynamic sports analytics query for level one QTC_B by 99,91%, using both the HLC and LLC as well as the non-compressed edit distance calculation, thus including the stopping mechanisms. This reduction was reported for a query ran on the football dataset, once using the non-optimized dynamic Python program, and once using the optimized version (by the inclusion of the above-mentioned heuristic). The former provides the desired solutions of the query (A), while the query ran on the optimized program gives the actual retrieved solutions (B). Using both, the performance of the optimized program can be evaluated. The precision of the heuristic describes the percentage of retrieved solutions that are correct ($\frac{A \cap B}{B}$), while the accuracy describes whether all of the desired solutions were found $\frac{A \cap B}{A}$). The usage of the heuristic, however, lead to a small decrease in accuracy for the example (football) dataset, because of the use of HLC in the first step. Although the precision of the heuristic stayed 100% (i.e. all the solutions were part of the ground truth) it can be seen as non-favourable, because of the significant decrease in the accuracy of the results (not all of the desired solutions were found). For this reason, the idea of this preliminary step for the proposed dynamic sports analytics method is not advised.

Additional optimization approaches

The optimization techniques mentioned in the previous sections are not the only options for optimizing the performance of the proposed methods for static and dynamic sports analytics. This section consists of some additional preliminary thoughts and approaches that were not implemented or subject of a detailed theoretical study, but are still worth mentioning.

For dynamic sports analytics, for example, it might be interesting to first compare fragments with a coarser temporal resolution. Since the number of compared fragments is often large, it can be feasible to only calculate the distance between the fragments with the original (finer) temporal resolution when they have proven to be similar at a coarser temporal resolution. Similarly, as a preliminary step, one could reduce the number of included MPOs that make up the QTC-matrix sequences (dynamic sports analytics) or the vectors that make up the QTC_S-matrices (static sports analytics). Only the fragments or moments that turn out to be very similar based on these reduced QTC-representations could thereafter be compared by using all of the desired MPOs.

When multiple queries with largely similar parameters are executed on the same dataset, it might be feasible to calculate the exhaustive QTC-representations for that dataset (*i.e.* including all MPOs and all timestamps), rather than sequentially re-calculating the QTC-representations of the requested timestamps and MPOs. Indeed, the latter implies that identical QTC-relations are calculated multiple times because they are not stored after every query. To enhance the performance in such a case, the exhaustive QTC-representation could be computed once, after which a selection is applied on this representation for each query that is executed.

From a coding perspective, the performance of the dynamic and static sports analytics methods could be enhanced by a parallelization of the code. Parallel computing is characterized by the division of one computational task into several parts that are calculated simultaneously on multiple computer cores (Almasi & Gottlieb, 1989). By calculating and comparing the QTC-representations of different fragments or moments simultaneously on multiple cores, the performance of both methods could be improved significantly.

References

Almasi, G.S., Gottlieb, A. (1989). Highly parallel computing. Redwood City, CA, U.S.A.: Benjamin-Cummings Publishing Co.

Robinson, A.H., Cherry, C. (1967). Results of a prototype television bandwidth compression scheme. In Proceedings of the IEEE (pp. 356-364). Piscataway, NJ, U.S.A.: Institute of Electrical and Electronics Engineers

Standaert, H. (2018). *Analysis of large amounts of motion trajectories.* Master thesis dissertation, Ghent University.

Appendix 3
Use Case 1: Duo-trio test

Dear participant,

My name is Jasper Beernaerts, I am a PhD Student at the Department of Geography, doing research on pattern recognition in football and other sports and prof. Lenoir is one of my promoters. Currently I am analysing how to detect similar actions between two footballers during a real professional football match. For this, I analyse the running patterns of the football players, thus only taking into account positional data.

I would like to present you a small test where you would have to indicate, only based on the running patterns of two players, which fragments are more similar. The test consists of 18 questions containing a reference fragment of 20 seconds and two sample fragments of 20 seconds. You need to indicate which of both sample fragments most resembles the reference fragment, thereby only basing yourself on the running patterns of the players. You will get a template where you need to fill in your name and a 'A' or 'B' for every question. During the test, you are not allowed to communicate with your colleagues and you are not allowed to change answers of previous questions. Answering each question is mandatory.

Thank you very much for your participation!

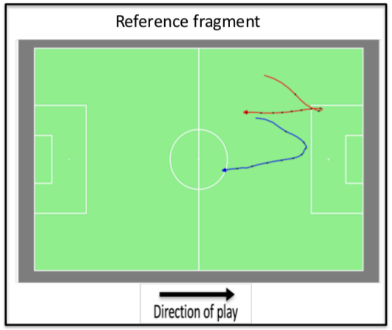
Questionnaire indicating similar football patterns

Please enter full screen mode by starting the slide show

This questionnaire contains **18 questions**

Each question consists of:

- <u>one reference fragment</u> (identical for all questions)
- <u>two sample fragments</u> (differing between questions)



The <u>reference fragment</u> displays the movements of two players of the same team (in red and blue) on a football field during 20 seconds.

The <u>sample fragments</u> display the movements of the same two players as in the reference fragment, during different periods of 20 seconds.

Small black dots show the positions of the players every 2 seconds

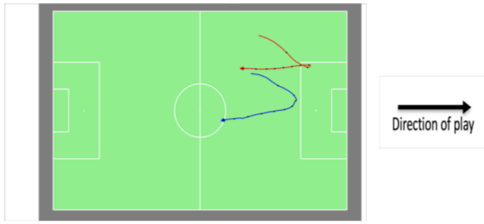
Please note down for each question which sample fragment (A or B) is most similar to the reference fragment (blank answers are not allowed).

-Once answered, it is not allowed to return to a previous question.

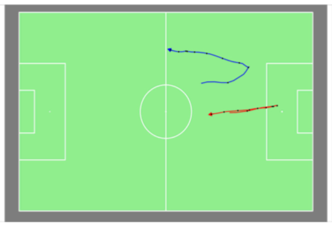
-No communication is allowed

Question 1

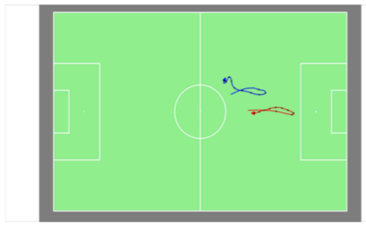
Reference fragment



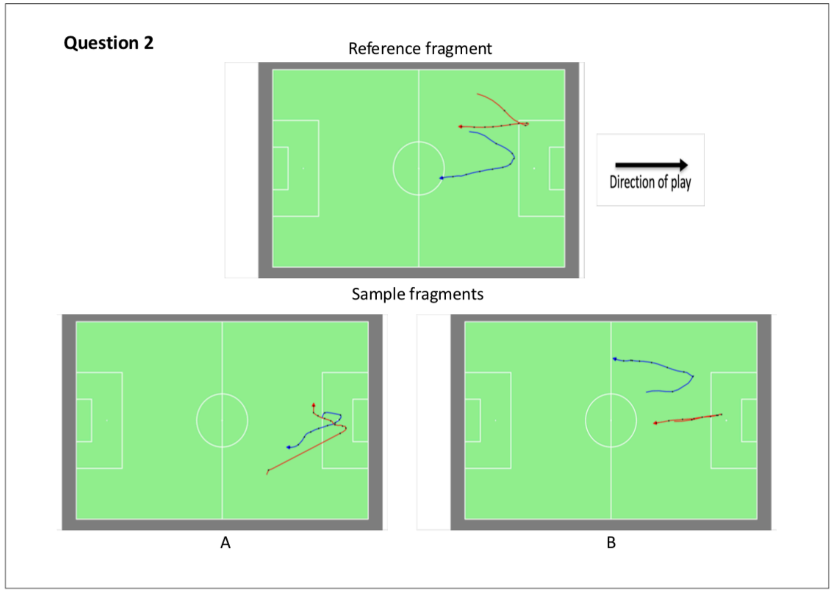
Sample fragments

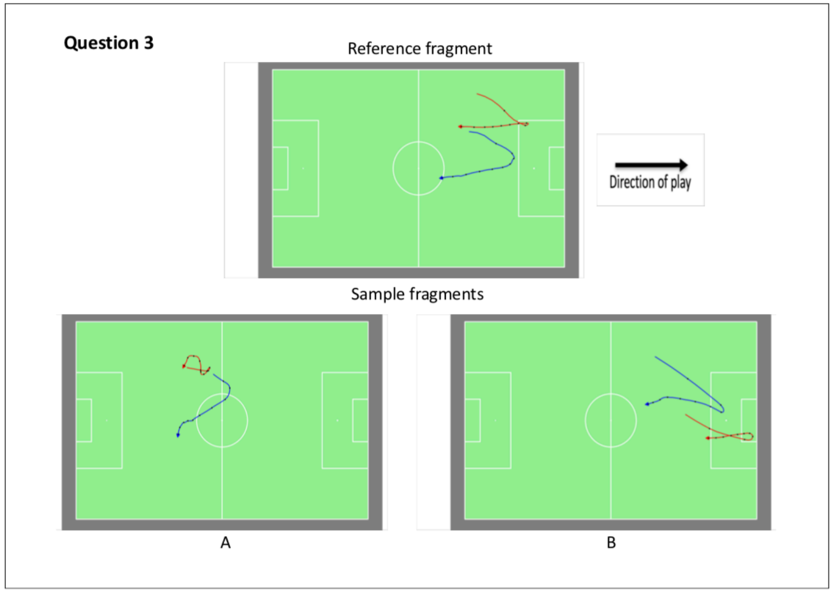


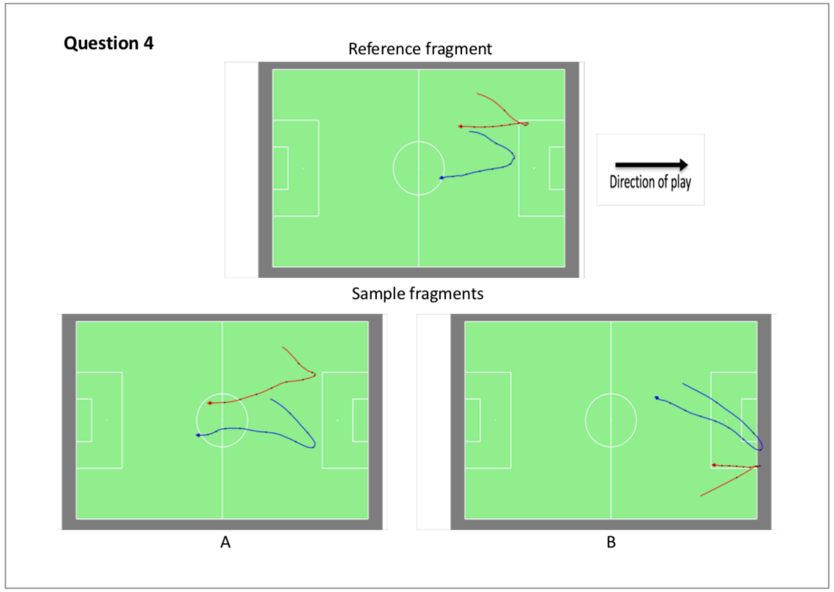
A

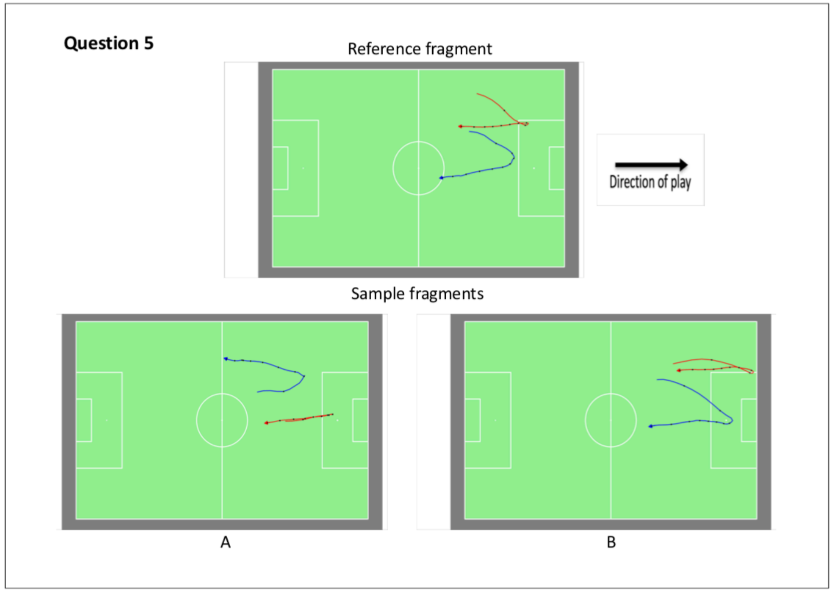


B


Question 2
Reference fragment
Direction of play
Sample fragments
A
B

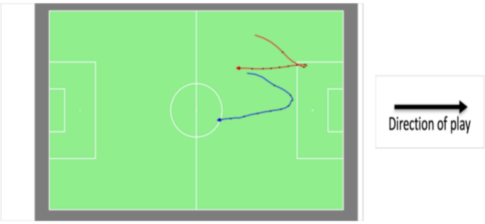

Question 3
Reference fragment
Direction of play
Sample fragments
A
B


Question 4
Reference fragment
Direction of play
Sample fragments
A
B

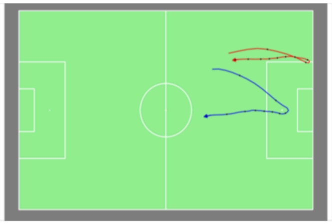

Question 5
Reference fragment
Direction of play
Sample fragments
A
B

Question 6

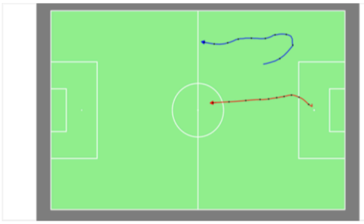
Reference fragment



Sample fragments



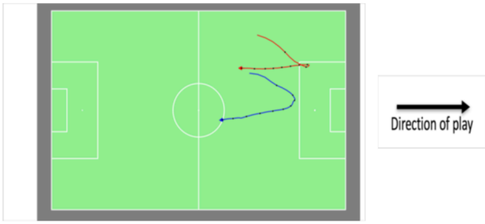
A



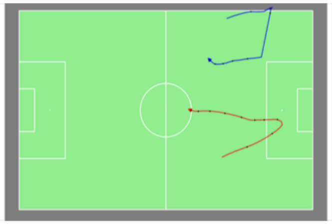
B

Question 7

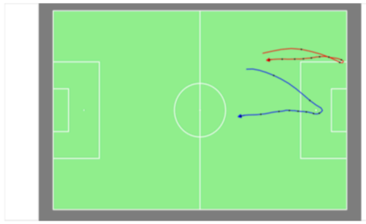
Reference fragment



Sample fragments



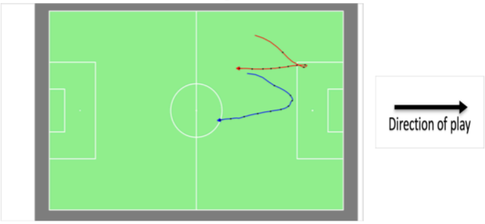
A



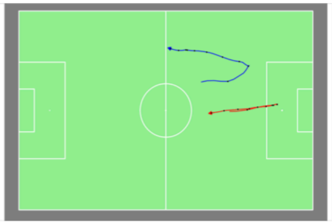
B

Question 8

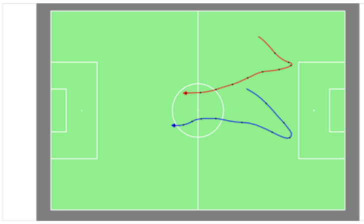
Reference fragment



Sample fragments



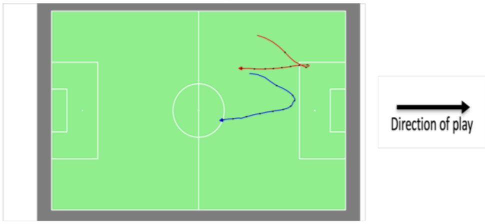
A



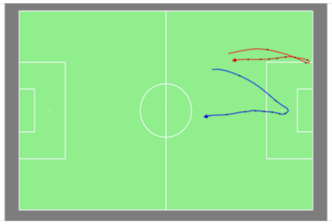
B

Question 9

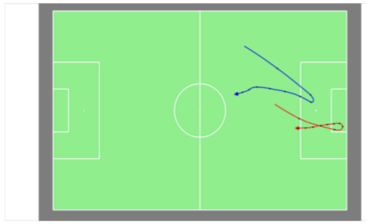
Reference fragment



Sample fragments



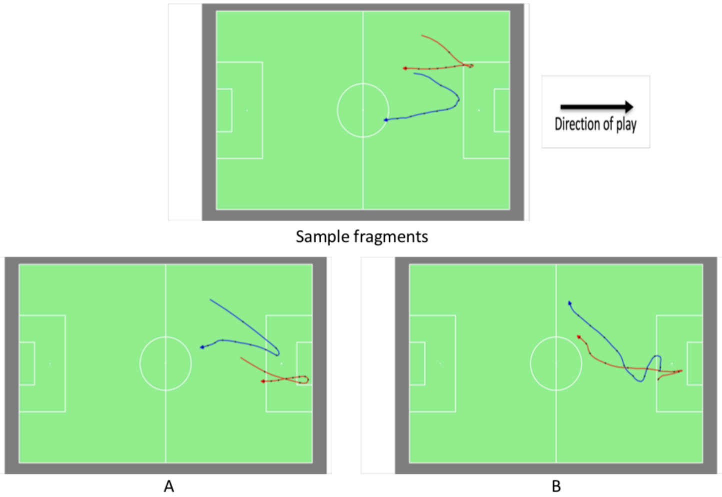
A



B

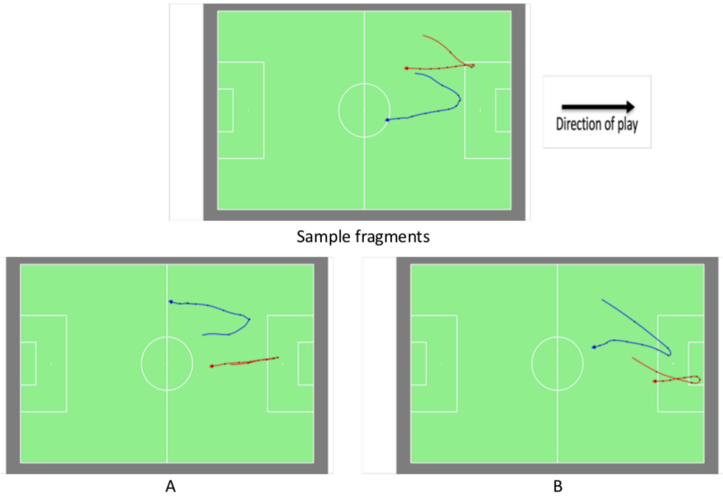
Question 10

Reference fragment



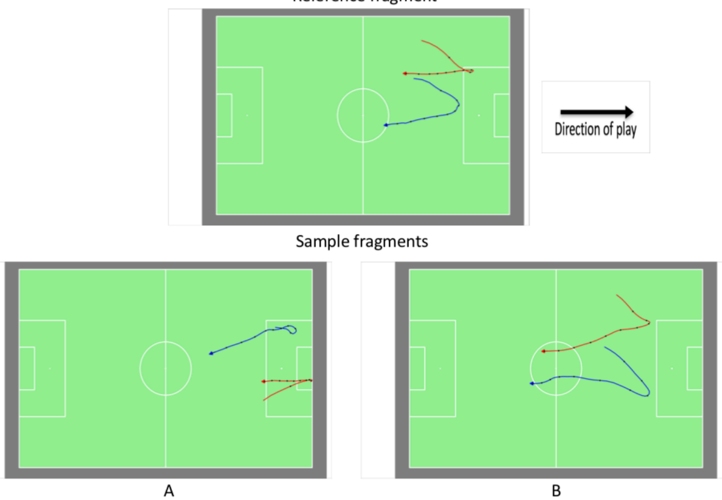
Question 11

Reference fragment



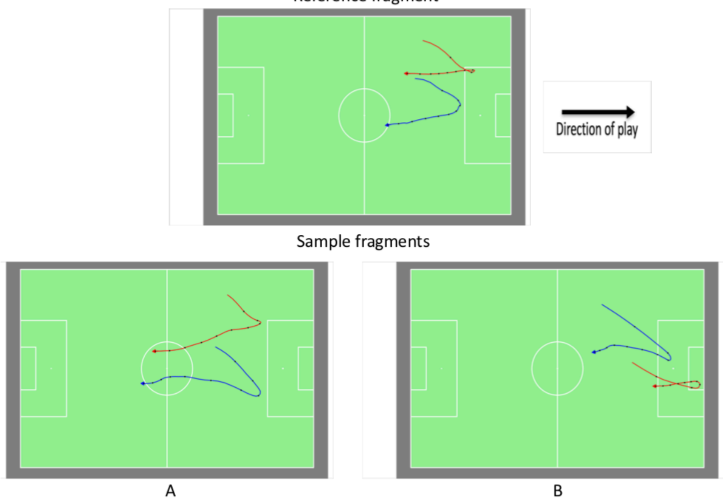
Question 12

Reference fragment

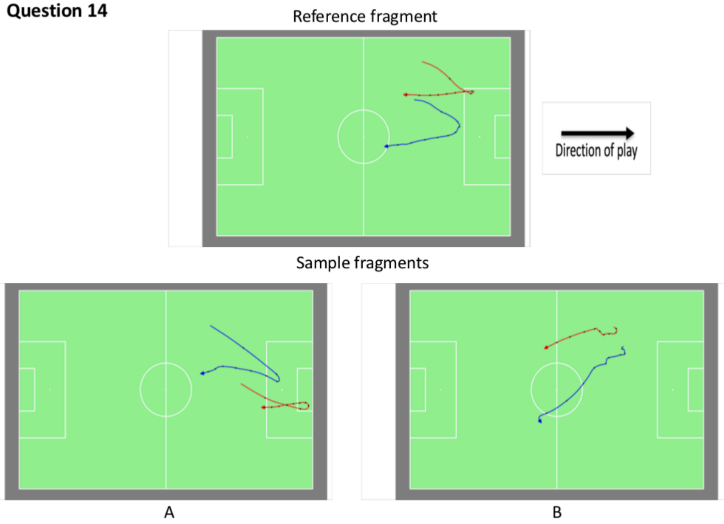


Question 13

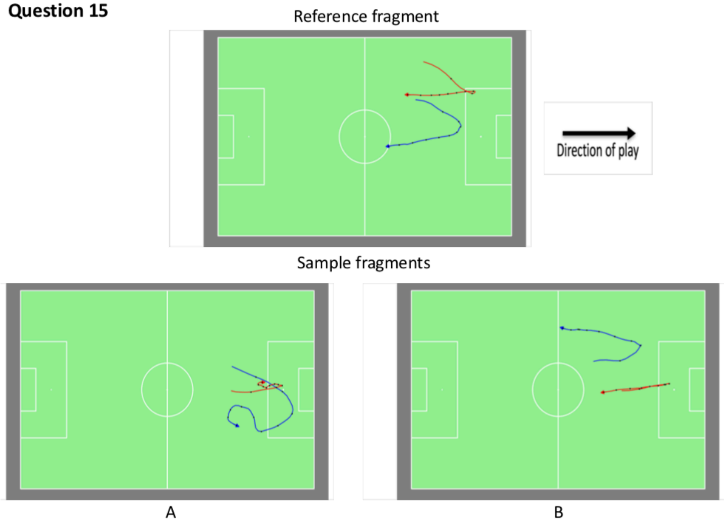
Reference fragment



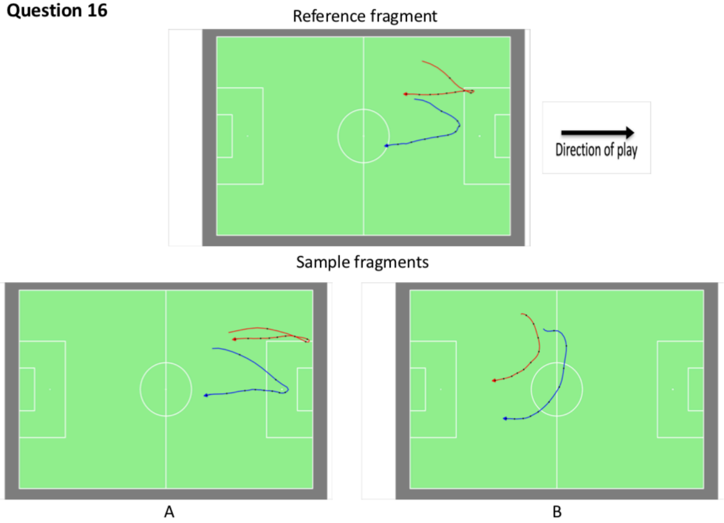
Question 14



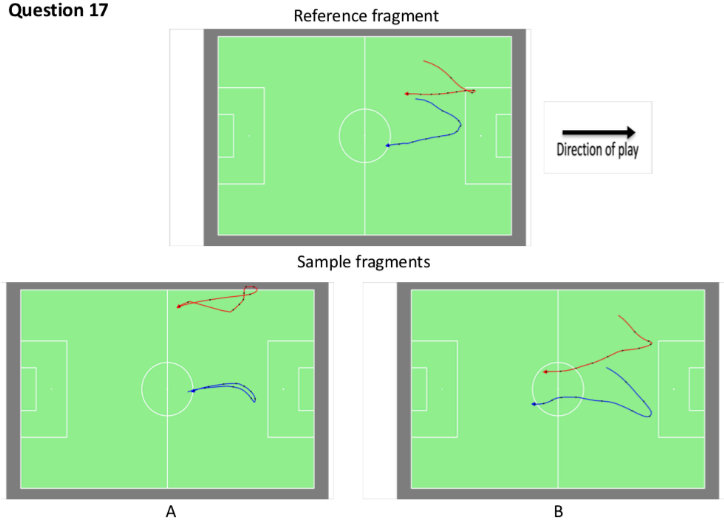
Question 15



Question 16

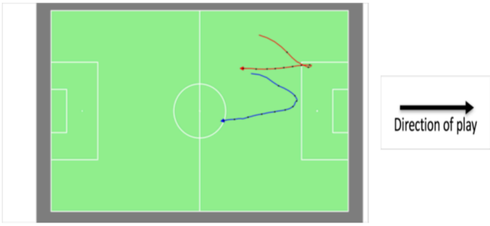


Question 17

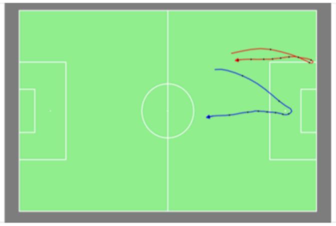


Question 18

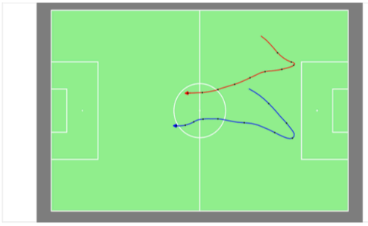
Reference fragment



Sample fragments



A



B

Any remarks?

- **How did you decide which fragment was most similar?**
- **Was the test difficult?**
- **Suggestions?**
- **...**

Thank you for your participation

Question number	Question Type	Rank number fragment 1	Rank number fragment 2	Correct answer	Participant number																		
					1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	Aub_C	3	6	**A**	B	A	A	A	A	A	A	A	A	A	A	A	A	A	B	B	B	A	A
2	Aub_C	10	3	**B**	A	B	B	B	B	A	B	B	B	B	B	B	B	B	A	B	A	B	A
3	Aub_C	8	4	**B**	A	B	B	B	B	B	B	B	B	B	B	B	B	B	A	B	B	B	A
4	Aub_C	1	9	**A**	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
5	A_B	3	2	**B**	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	A	B	B
6	Aub_C	2	13	**A**	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
7	Aub_C	16	2	**B**	B	B	B	B	B	B	B	B	B	B	B	B	B	A	B	B	A	B	B
8	A_B	3	1	**B**	B	B	B	B	B	B	B	B	B	B	B	B	B	A	B	B	A	B	B
9	A_B	2	4	**A**	A	A	A	A	A	A	A	A	A	B	A	A	A	B	A	A	B	A	A
10	Aub_C	4	12	**A**	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	B	B	A	B
11	B_B	3	4	**A**	A	B	B	A	A	B	B	B	B	B	A	A	A	B	B	B	B	B	B
12	Aub_C	5	1	**B**	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B
13	A_B	1	4	**A**	A	A	B	B	A	A	A	A	A	A	A	A	A	A	B	A	B	A	A
14	Aub_C	4	15	**A**	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	B
15	Aub_C	11	3	**B**	B	B	A	A	B	B	B	B	A	B	B	B	B	B	B	A	A	B	A
16	Aub_C	2	7	**A**	A	A	A	B	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
17	Aub_C	14	1	**B**	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B
18	A_A	2	1	**B**	A	A	A	A	B	A	A	A	A	B	B	B	A	A	A	B	B	A	A

Question number	Question Type	Rank number fragment 1	Rank number fragment 2	Correct answer	Participant number																	
					20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37
1	Aub_C	3	6	**A**	A	A	A	B	A	A	A	A	B	A	A	B	A	B	A	B	B	B
2	Aub_C	10	3	**B**	B	B	A	B	B	B	B	B	A	B	A	B	B	B	B	B	A	B
3	Aub_C	8	4	**B**	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	A	B	A
4	Aub_C	1	9	**A**	A	A	A	A	A	A	B	A	A	A	A	B	A	A	A	A	A	A
5	A_B	3	2	**B**	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B
6	Aub_C	2	13	**A**	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
7	Aub_C	16	2	**B**	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	A
8	A_B	3	1	**B**	A	A	B	B	B	B	A	B	A	B	B	B	B	B	B	A	B	B
9	A_B	2	4	**A**	A	A	A	B	A	B	A	A	A	A	A	A	A	A	A	A	A	B
10	Aub_C	4	12	**A**	A	A	B	A	A	A	A	A	B	A	A	A	A	A	A	A	A	A
11	B_B	3	4	**A**	A	B	B	B	B	B	B	B	A	A	B	B	B	B	B	B	A	B
12	Aub_C	5	1	**B**	B	B	B	B	B	B	A	B	B	B	B	A	B	B	B	A	B	B
13	A_B	1	4	**A**	B	B	A	B	A	A	B	A	A	A	B	B	B	A	A	B	A	A
14	Aub_C	4	15	**A**	A	A	B	A	A	A	A	B	A	A	A	A	A	A	B	A	A	B
15	Aub_C	11	3	**B**	B	B	A	B	A	B	B	B	A	B	A	B	B	A	B	B	B	B
16	Aub_C	2	7	**A**	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	B
17	Aub_C	14	1	**B**	B	A	B	A	B	B	B	B	B	B	B	A	B	B	B	A	B	B
18	A_A	2	1	**B**	A	A	B	A	A	B	A	A	B	B	A	A	A	A	A	B	B	A

Appendix 4
Use case 2: Classification errors for all marker combinations

2 marker combinations

	k	ankle, elbow	ankle, knee	ankle, shoulder	ankle, toe	ankle, troch	ankle, wrist	elbow, toe	knee, elbow	knee, shoulder	knee, toe	knee, troch
Error at step level	1	0.4444	0.3611	0.4306	0.4167	0.2917	0.3750	0.3611	0.4028	0.5417	0.6250	0.4167
	3	0.4028	0.4028	0.5278	0.4861	0.3194	0.3472	0.3889	0.4444	0.4722	0.6250	0.4167
	5	0.3750	0.4028	0.5139	0.4028	0.3750	0.3611	0.3333	0.4028	0.5278	0.5694	0.4861
	7	0.3750	0.3889	0.4722	0.4167	0.4028	0.3611	0.3889	0.4306	0.4861	0.5556	0.4583
	9	0.3750	0.4028	0.4306	0.4444	0.4028	0.3611	0.4028	0.4028	0.3750	0.5556	0.4861
	11	0.3889	0.4306	0.4028	0.4306	0.3472	0.3889	0.3889	0.3889	0.4306	0.5417	0.4306
	13	0.4167	0.4583	0.4583	0.3611	0.3889	0.4167	0.3750	0.4167	0.4722	0.5139	0.4167
Error at individual level	1	0.3889	0.3333	0.3889	0.3333	0.3333	0.3889	0.2778	0.3333	0.5556	0.6667	0.4444
	3	0.4444	0.2778	0.5556	0.3333	0.3333	0.3889	0.3333	0.3333	0.5556	0.5556	0.5000
	5	0.3333	0.3333	0.5000	0.2778	0.3333	0.2778	0.3333	0.3333	0.5000	0.5556	0.4444
	7	0.3333	0.4444	0.4444	0.3889	0.3333	0.3333	0.3889	0.3333	0.5000	0.5000	0.4444
	9	0.3333	0.4444	0.3889	0.2222	0.3889	0.2778	0.3333	0.3333	0.3889	0.5000	0.4444
	11	0.3889	0.4444	0.5000	0.3333	0.3333	0.3333	0.3889	0.3333	0.4444	0.5000	0.4444
	13	0.3889	0.5000	0.4444	0.2222	0.3333	0.3333	0.2778	0.3333	0.4444	0.3889	0.3889

	k	knee, wrist	shoulder, elbow	shoulder, toe	troch, elbow	troch, shoulder	troch, toe	troch, wrist	wrist, elbow	wrist, shoulder	wrist, toe
Error at step level	1	0.4861	0.4583	0.3194	0.4028	0.4583	0.3750	0.4306	0.4306	0.4306	0.4583
	3	0.4306	0.4167	0.3750	0.3750	0.4444	0.3889	0.4167	0.4583	0.4167	0.4028
	5	0.4167	0.5417	0.4167	0.3889	0.4028	0.3611	0.4028	0.4167	0.3472	0.4028
	7	0.3472	0.5000	0.4167	0.4306	0.4583	0.3611	0.3611	0.4306	0.4306	0.3611
	9	0.3611	0.4861	0.4028	0.3750	0.4583	0.3750	0.3333	0.4861	0.4722	0.3750
	11	0.3333	0.4722	0.3889	0.3889	0.4444	0.3611	0.3472	0.5417	0.4722	0.4444
	13	0.3333	0.4028	0.4028	0.3889	0.4444	0.4167	0.3472	0.5972	0.4444	0.4167
Error at individual level	1	0.5556	0.4444	0.2778	0.5000	0.4444	0.3889	0.5000	0.2778	0.3889	0.6111
	3	0.5000	0.3333	0.2778	0.3333	0.3889	0.3889	0.3889	0.4444	0.2778	0.5556
	5	0.4444	0.4444	0.3333	0.4444	0.3889	0.3889	0.3333	0.4444	0.2222	0.4444
	7	0.3333	0.3889	0.3333	0.3889	0.3333	0.2778	0.3333	0.3889	0.3889	0.2222
	9	0.3333	0.3333	0.4444	0.4444	0.3333	0.3333	0.3333	0.5556	0.3889	0.3333
	11	0.2778	0.3889	0.3889	0.3889	0.3333	0.3333	0.3333	0.5556	0.3889	0.5000
	13	0.2778	0.4444	0.3889	0.3889	0.3333	0.3333	0.3333	0.6111	0.4444	0.4444

3 marker combinations

	k	ankle, elbow, wrist	ankle, knee, elbow	ankle, knee, shoulder	ankle, knee, troch	ankle, knee, wrist	ankle, shoulder, elbow	ankle, shoulder, wrist	ankle, troch, elbow	ankle, troch, shoulder	ankle, troch, wrist	knee, elbow, toe	knee, elbow, wrist
Error at step level	1	0.4722	0.4722	0.5278	0.4583	0.4444	0.3889	0.4306	0.4028	0.4583	0.3750	0.5417	0.4722
	3	0.4583	0.4167	0.4861	0.4167	0.4583	0.4583	0.3889	0.3889	0.4444	0.3333	0.5278	0.5278
	5	0.4028	0.5139	0.5000	0.4306	0.3750	0.4306	0.3611	0.4444	0.4167	0.2778	0.4722	0.4444
	7	0.4028	0.4444	0.4861	0.4028	0.3472	0.4583	0.4167	0.4167	0.3750	0.3194	0.4583	0.4861
	9	0.4444	0.3333	0.4306	0.4583	0.3889	0.4306	0.3889	0.3611	0.3889	0.2778	0.4583	0.4722
	11	0.4444	0.3333	0.4444	0.4167	0.3750	0.3889	0.3611	0.3889	0.4028	0.2917	0.4444	0.4306
	13	0.4583	0.4167	0.4583	0.4167	0.3611	0.4167	0.3750	0.3750	0.4167	0.3333	0.4583	0.4583
Error at individual level	1	0.4444	0.5000	0.5556	0.4444	0.4444	0.5000	0.5000	0.3889	0.5000	0.3889	0.7222	0.5556
	3	0.4444	0.3333	0.5000	0.3333	0.5556	0.3333	0.4444	0.3333	0.3333	0.3333	0.5000	0.3889
	5	0.3889	0.3333	0.4444	0.2778	0.3889	0.4444	0.3889	0.2222	0.3333	0.3333	0.3889	0.3889
	7	0.3889	0.2778	0.4444	0.3333	0.3333	0.3333	0.4444	0.3889	0.2222	0.2778	0.3889	0.4444
	9	0.4444	0.2778	0.3889	0.2778	0.2778	0.3333	0.3333	0.3333	0.2778	0.2778	0.3889	0.5556
	11	0.4444	0.3333	0.3889	0.2778	0.2778	0.3333	0.3889	0.3889	0.2778	0.3333	0.3889	0.5000
	13	0.4444	0.3333	0.5000	0.3333	0.3333	0.3333	0.3333	0.3889	0.2778	0.3333	0.3333	0.5556

	k	knee, shoulder, elbow	knee, shoulder, wrist	knee, troch, elbow	knee, troch, shoulder	shoulder, elbow, wrist	toe, ankle, elbow	toe, ankle, knee	toe, ankle, shoulder	toe, ankle, troch	toe, ankle, wrist	toe, elbow, wrist
Error at step level	1	0.5000	0.4722	0.3889	0.4028	0.4444	0.3889	0.4444	0.4167	0.2917	0.3333	0.3889
	3	0.4583	0.4306	0.4028	0.4028	0.4861	0.4444	0.5278	0.5000	0.3333	0.3889	0.5139
	5	0.4583	0.4028	0.4583	0.3750	0.4861	0.4028	0.4306	0.4722	0.3472	0.3472	0.5417
	7	0.4444	0.3333	0.4861	0.4028	0.4861	0.3750	0.5000	0.4306	0.3750	0.3750	0.4861
	9	0.4861	0.3889	0.4722	0.4167	0.4861	0.4028	0.5139	0.4583	0.4306	0.4167	0.4722
	11	0.4444	0.3889	0.4861	0.4028	0.5139	0.4306	0.5139	0.4444	0.4306	0.4306	0.4444
	13	0.4444	0.3889	0.4583	0.4167	0.5139	0.3750	0.4861	0.4583	0.4444	0.4444	0.4306
Error at individual level	1	0.5556	0.6667	0.3333	0.3889	0.5556	0.3333	0.5000	0.5000	0.2778	0.2222	0.3889
	3	0.3333	0.4444	0.4444	0.3889	0.5556	0.3333	0.4444	0.3889	0.1667	0.4444	0.5000
	5	0.3889	0.2778	0.4444	0.2778	0.4444	0.3333	0.3889	0.3333	0.3333	0.3889	0.5556
	7	0.3889	0.3333	0.3889	0.2778	0.5556	0.3333	0.5000	0.3333	0.3333	0.3333	0.5556
	9	0.4444	0.3333	0.4444	0.3889	0.5000	0.2778	0.3889	0.3333	0.3889	0.3889	0.5000
	11	0.4444	0.3889	0.4444	0.3333	0.4444	0.2778	0.3889	0.4444	0.3889	0.4444	0.5556
	13	0.5556	0.3333	0.4444	0.3333	0.5000	0.2222	0.3333	0.3889	0.3889	0.3889	0.3889

	k	toe, knee, shoulder	toe, knee, troch	toe, knee, wrist	toe, shoulder, elbow	toe, shoulder, wrist	toe, troch, elbow	toe, troch, shoulder	toe, troch, wrist	troch, elbow, wrist	troch, shoulder, elbow	troch, shoulder, wrist
Error at step level	1	0.4444	0.4444	0.4444	0.4167	0.4444	0.4444	0.3472	0.3889	0.3889	0.3750	0.3611
	3	0.5833	0.4444	0.5000	0.4167	0.5278	0.3750	0.3333	0.3333	0.3750	0.3056	0.3194
	5	0.4861	0.4444	0.4722	0.3750	0.4444	0.4167	0.3472	0.3194	0.3889	0.3611	0.3056
	7	0.4861	0.5000	0.4722	0.4028	0.3889	0.4444	0.3472	0.3194	0.3889	0.3333	0.3472
	9	0.5139	0.4167	0.4722	0.3889	0.4028	0.4306	0.3333	0.3056	0.4028	0.3889	0.3194
	11	0.4444	0.3889	0.4444	0.4028	0.4028	0.4583	0.3194	0.3333	0.4167	0.4167	0.3194
	13	0.4444	0.4167	0.4444	0.3750	0.4306	0.4444	0.3472	0.3333	0.4167	0.4583	0.3611
Error at individual level	1	0.5000	0.3333	0.5000	0.3333	0.5000	0.5000	0.4444	0.3333	0.5000	0.2778	0.3889
	3	0.5556	0.3889	0.6111	0.3333	0.5556	0.3333	0.2222	0.3889	0.3889	0.3333	0.2778
	5	0.5000	0.4444	0.5556	0.2778	0.3889	0.4444	0.2222	0.3889	0.4444	0.3333	0.3333
	7	0.4444	0.4444	0.5556	0.2778	0.2778	0.3889	0.2222	0.2778	0.3333	0.3333	0.2778
	9	0.4444	0.4444	0.5556	0.3333	0.3333	0.3333	0.2222	0.2778	0.5000	0.3333	0.3333
	11	0.4444	0.3333	0.5556	0.2778	0.4444	0.3889	0.2222	0.3889	0.4444	0.3889	0.2778
	13	0.4444	0.2778	0.5556	0.3333	0.4444	0.3333	0.2778	0.3333	0.5000	0.4444	0.3333

4 marker combinations

	k	ankle, knee, elbow, wrist	ankle, knee, shoulder, elbow	ankle, knee, shoulder, wrist	ankle, knee, troch, elbow	ankle, knee, troch, shoulder	ankle, knee, troch, wrist	ankle, shoulder, elbow, wrist	ankle, troch, elbow, wrist	ankle, troch, shoulder, elbow	ankle, troch, shoulder, toe	ankle, troch, shoulder, wrist	knee, shoulder, elbow, wrist
Error at step level	1	0.5139	0.5139	0.4583	0.5000	0.4306	0.3611	0.4583	0.4444	0.4167	0.3750	0.3472	0.5417
	3	0.4444	0.5139	0.4306	0.4861	0.4722	0.3472	0.4583	0.3472	0.4306	0.3333	0.3611	0.5000
	5	0.4583	0.4722	0.3889	0.4306	0.4306	0.3611	0.4722	0.3750	0.3611	0.3056	0.3333	0.4861
	7	0.4444	0.4306	0.3611	0.4444	0.3611	0.3056	0.4306	0.3472	0.3611	0.3750	0.3194	0.4861
	9	0.4444	0.5000	0.3889	0.4306	0.4167	0.2917	0.4306	0.3194	0.4028	0.3889	0.3472	0.4444
	11	0.4583	0.4583	0.4167	0.4583	0.4028	0.3472	0.4583	0.3333	0.4167	0.3889	0.3333	0.4444
	13	0.4028	0.4306	0.4444	0.4583	0.4306	0.3472	0.3889	0.3750	0.4167	0.3750	0.3611	0.4583
Error at individual level	1	0.6111	0.5000	0.6111	0.4444	0.4444	0.3889	0.6111	0.5000	0.4444	0.5000	0.4444	0.7222
	3	0.3889	0.4444	0.3889	0.4444	0.4444	0.2778	0.4444	0.2778	0.3333	0.1667	0.3333	0.4444
	5	0.4444	0.3333	0.2778	0.3889	0.3333	0.2778	0.4444	0.2778	0.2778	0.2222	0.2778	0.4444
	7	0.3889	0.2778	0.3889	0.3889	0.2778	0.2778	0.4444	0.3889	0.3333	0.2222	0.2778	0.3889
	9	0.3889	0.3889	0.3889	0.3889	0.3333	0.2778	0.3889	0.3333	0.3333	0.2222	0.3889	0.4444
	11	0.3889	0.3889	0.3333	0.3333	0.2778	0.3333	0.3889	0.3889	0.3333	0.2778	0.3333	0.3889
	13	0.3889	0.3889	0.3889	0.3333	0.2222	0.3889	0.3333	0.3889	0.3333	0.2222	0.3333	0.3333

	k	knee, troch, elbow, wrist	knee, troch, shoulder, elbow	knee, troch, shoulder, wrist	toe, ankle, elbow, wrist	toe, ankle, knee, elbow	toe, ankle, knee, shoulder	toe, ankle, knee, troch	toe, ankle, knee, wrist	toe, ankle, shoulder, elbow	toe, ankle, shoulder, wrist	toe, ankle, troch, elbow	toe, ankle, troch, wrist
Error at step level	1	0.4306	0.3611	0.3472	0.4167	0.4722	0.4861	0.3889	0.4306	0.3889	0.4583	0.4444	0.3472
	3	0.4306	0.4028	0.3611	0.4722	0.4861	0.5139	0.4167	0.4167	0.4167	0.4028	0.4306	0.3056
	5	0.3611	0.4722	0.3472	0.4028	0.4583	0.4722	0.4722	0.4306	0.4167	0.3750	0.3750	0.2778
	7	0.4167	0.4583	0.3611	0.3889	0.4306	0.4722	0.4861	0.4028	0.3750	0.3750	0.3750	0.3194
	9	0.3889	0.4444	0.3750	0.4444	0.4028	0.5139	0.3750	0.4444	0.4306	0.4167	0.3194	0.2639
	11	0.4306	0.4861	0.3611	0.4444	0.4444	0.5000	0.4028	0.4306	0.3889	0.4028	0.3611	0.2639
	13	0.4167	0.4167	0.4167	0.4306	0.4722	0.5000	0.3750	0.4306	0.3889	0.4306	0.3472	0.2778
Error at individual level	1	0.5000	0.2778	0.2778	0.4444	0.4444	0.5000	0.3889	0.3889	0.3333	0.6111	0.4444	0.3889
	3	0.3333	0.3889	0.2778	0.4444	0.4444	0.5556	0.2778	0.3889	0.2222	0.3889	0.2778	0.2778
	5	0.2778	0.3889	0.2778	0.3889	0.3333	0.3889	0.3333	0.4444	0.3333	0.3889	0.2778	0.2778
	7	0.3889	0.3889	0.2222	0.3889	0.3333	0.4444	0.2778	0.5556	0.2778	0.3889	0.2778	0.2778
	9	0.3333	0.3889	0.3333	0.3889	0.3333	0.4444	0.2778	0.4444	0.3333	0.3333	0.2778	0.3889
	11	0.3889	0.3889	0.3333	0.3889	0.3333	0.4444	0.2778	0.4444	0.2778	0.3889	0.2778	0.3889
	13	0.4444	0.3889	0.3889	0.3333	0.3889	0.5000	0.2778	0.3889	0.3889	0.3889	0.2778	0.3333

	k	toe, knee, elbow, wrist	toe, knee, shoulder, elbow	toe, knee, shoulder, wrist	toe, knee, troch, elbow	toe, knee, troch, shoulder	toe, knee, troch, wrist	toe, shoulder, elbow, wrist	toe, troch, elbow, wrist	toe, troch, shoulder, elbow	toe, troch, shoulder, wrist	troch, shoulder, elbow, wrist
Error at step level	1	0.4861	0.5000	0.5556	0.4722	0.4167	0.3889	0.4722	0.4306	0.3611	0.3750	0.4444
	3	0.5139	0.5278	0.5000	0.4444	0.3750	0.3472	0.4306	0.4028	0.3611	0.3194	0.3750
	5	0.4861	0.5139	0.4722	0.4861	0.4028	0.3472	0.4167	0.3750	0.4306	0.3611	0.3472
	7	0.4444	0.4444	0.5000	0.4583	0.4306	0.3750	0.4444	0.4028	0.4167	0.3194	0.3611
	9	0.4444	0.4722	0.4444	0.4444	0.4722	0.3194	0.4306	0.3889	0.3750	0.3611	0.3750
	11	0.4861	0.4722	0.3889	0.4583	0.4306	0.3194	0.4444	0.4028	0.4444	0.3611	0.4028
	13	0.4861	0.4722	0.4028	0.4306	0.4167	0.3472	0.3889	0.4028	0.3889	0.3889	0.4861
Error at individual level	1	0.5556	0.5556	0.6667	0.5000	0.4444	0.3889	0.5000	0.5556	0.2778	0.4444	0.5000
	3	0.5556	0.4444	0.5000	0.4444	0.2222	0.3333	0.4444	0.3889	0.2222	0.3333	0.3889
	5	0.5556	0.3333	0.4444	0.3889	0.3889	0.3889	0.3889	0.2778	0.2778	0.3333	0.2778
	7	0.5000	0.3333	0.4444	0.3889	0.3333	0.3889	0.3333	0.2778	0.3889	0.3333	0.3333
	9	0.5000	0.3889	0.4444	0.3889	0.3889	0.2778	0.3889	0.2778	0.2778	0.2778	0.3333
	11	0.5556	0.4444	0.4444	0.3889	0.3889	0.2778	0.5000	0.3333	0.3333	0.4444	0.3889
	13	0.5556	0.4444	0.3889	0.3889	0.3333	0.3333	0.4444	0.4444	0.3333	0.4444	0.3889

5 marker combinations

	k	*ankle, knee, shoulder, elbow, toe*	*ankle, knee, troch, elbow, toe*	*ankle,knee,troch,shoulder,elbow*	*ankle,knee,troch,shoulder,toe*	*ankle,knee,troch,wrist,elbow*	*ankle,knee,troch,wrist,shoulder*	*ankle,knee,troch,wrist,toe*	*ankle,knee,wrist,elbow,toe*	*ankle,knee,wrist,shoulder,elbow*	*ankle,knee,wrist,shoulder,toe*	*ankle,troch,shoulder,elbow,toe*
Error at step level	1	0.5000	0.4722	0.4861	0.4167	0.4583	0.3889	0.3611	0.4722	0.5000	0.5278	0.3889
	3	0.4583	0.4028	0.4444	0.3750	0.3750	0.3611	0.3611	0.4306	0.4028	0.4861	0.3889
	5	0.4444	0.4167	0.4722	0.4306	0.4028	0.3333	0.3472	0.5000	0.4861	0.4861	0.3889
	7	0.4583	0.4167	0.4583	0.4028	0.3194	0.3333	0.3333	0.4444	0.4167	0.4167	0.4167
	9	0.4444	0.3889	0.4444	0.4028	0.3750	0.3472	0.3333	0.4583	0.4444	0.4028	0.4028
	11	0.4583	0.3889	0.5139	0.4028	0.4028	0.3333	0.3750	0.5139	0.3750	0.4028	0.3611
	13	0.4167	0.4028	0.4306	0.3750	0.4306	0.3333	0.3611	0.4306	0.3889	0.4306	0.3750
Error at individual level	1	0.6111	0.4444	0.5000	0.4444	0.6111	0.3889	0.3889	0.5000	0.6111	0.6667	0.3333
	3	0.3889	0.2778	0.4444	0.2778	0.2778	0.2778	0.3889	0.3333	0.2778	0.4444	0.1667
	5	0.3333	0.3333	0.4444	0.2778	0.3333	0.2778	0.2778	0.4444	0.3889	0.5000	0.2222
	7	0.4444	0.3889	0.3889	0.3889	0.2778	0.3333	0.2778	0.5000	0.4444	0.5000	0.3333
	9	0.3333	0.3333	0.3333	0.3889	0.2778	0.2222	0.3889	0.3889	0.3889	0.3333	0.2778
	11	0.4444	0.3333	0.3333	0.3333	0.3889	0.2222	0.3333	0.4444	0.3889	0.3889	0.3333
	13	0.3889	0.3333	0.3333	0.3333	0.3889	0.2778	0.2778	0.3889	0.3333	0.4444	0.2778

	k	*ankle,troch,wrist,elbow,toe*	*ankle,troch,wrist,shoulder,elbow*	*ankle,troch,wrist,shoulder,toe*	*ankle,wrist,shoulder,elbow,toe*	*knee,troch,shoulder,elbow,toe*	*knee,troch,wrist,elbow,toe*	*knee,troch,wrist,shoulder,elbow*	*knee,troch,wrist,shoulder,toe*	*knee,wrist,shoulder,elbow,toe*	*troch,wrist,shoulder,elbow,toe*
Error at step level	1	0.4861	0.4306	0.3472	0.4583	0.4444	0.4722	0.5139	0.4444	0.5278	0.4583
	3	0.3889	0.4028	0.3472	0.4444	0.3889	0.4444	0.4028	0.3889	0.5000	0.4167
	5	0.4306	0.4444	0.3194	0.4028	0.4444	0.4722	0.4306	0.3750	0.5139	0.4306
	7	0.3333	0.3611	0.3194	0.3750	0.4722	0.4306	0.3472	0.3472	0.4861	0.3750
	9	0.4028	0.4028	0.3611	0.4167	0.4306	0.4861	0.4306	0.3472	0.4583	0.3889
	11	0.3611	0.3889	0.3333	0.4028	0.4167	0.4306	0.4167	0.3611	0.4167	0.3889
	13	0.3333	0.3750	0.3611	0.4306	0.3889	0.4583	0.4444	0.3889	0.4028	0.4167
Error at individual level	1	0.5556	0.4444	0.3333	0.5556	0.5000	0.5556	0.5000	0.5000	0.6111	0.5000
	3	0.3889	0.3889	0.2778	0.4444	0.3333	0.4444	0.2778	0.3889	0.4444	0.3889
	5	0.3889	0.3889	0.2222	0.4444	0.3889	0.3889	0.2778	0.2222	0.5556	0.3889
	7	0.2778	0.3889	0.2778	0.3889	0.3889	0.4444	0.3333	0.2222	0.3889	0.3333
	9	0.3333	0.3889	0.2222	0.4444	0.3889	0.5556	0.3889	0.3333	0.3889	0.3889
	11	0.2778	0.3889	0.3889	0.4444	0.3889	0.5556	0.3889	0.3889	0.4444	0.3889
	13	0.3333	0.2778	0.3333	0.4444	0.3333	0.4444	0.3889	0.2778	0.4444	0.3889

6 and 7 marker combinations

6 markers

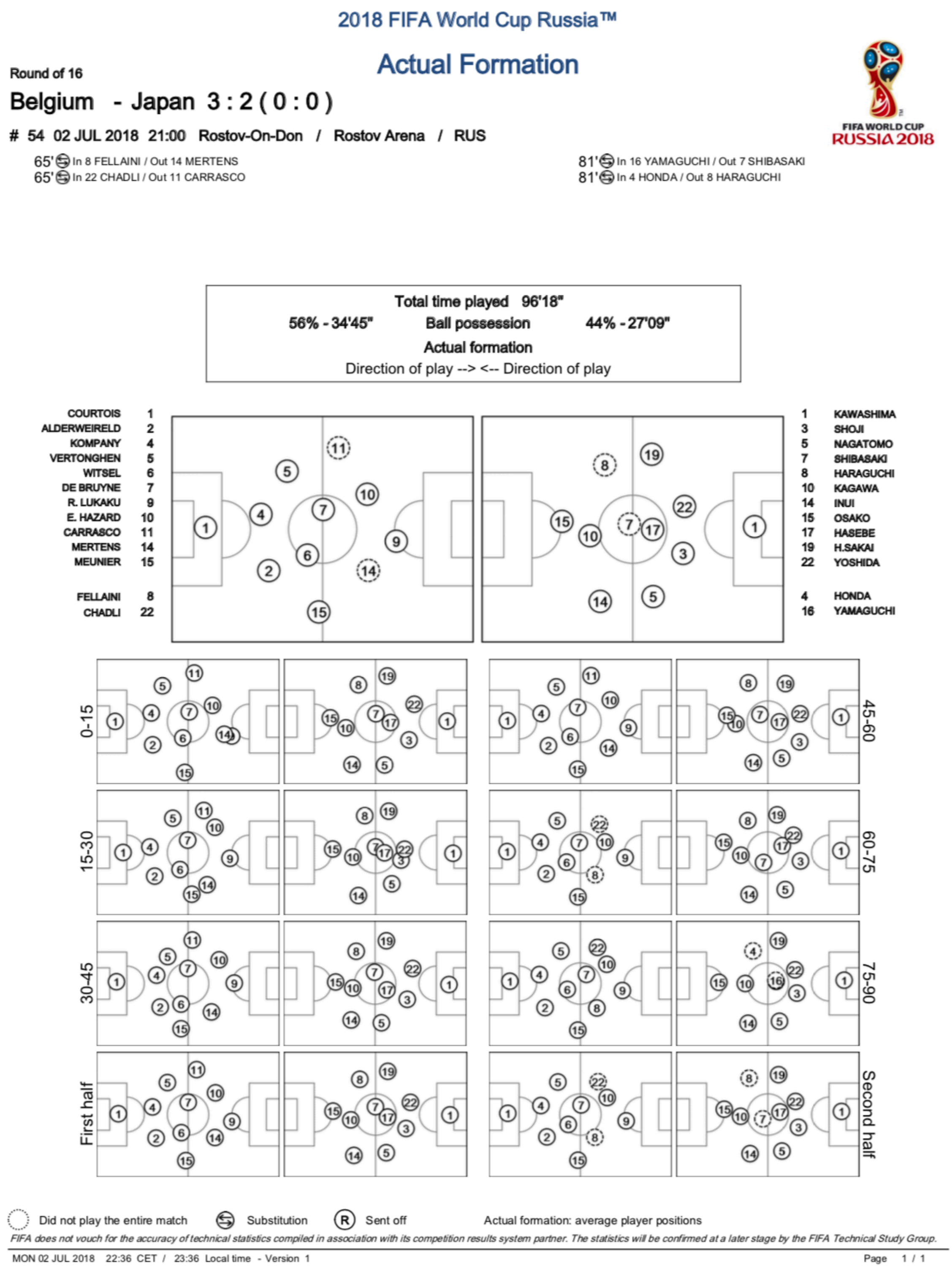
	k	ankle,knee,troch,shoulder,elbow,toe	ankle,knee,troch,wrist,elbow,toe	ankle,knee,troch,wrist,shoulder,elbow	ankle,knee,troch,wrist,shoulder,toe	ankle,knee,wrist,shoulder,elbow,toe	ankle,troch,wrist,shoulder,elbow,toe	knee,troch,wrist,shoulder,elbow,toe
Error at step level	1	0.4722	0.4583	0.5000	0.4583	0.5139	0.4444	0.5000
	3	0.3889	0.4306	0.4306	0.3889	0.4861	0.4444	0.4444
	5	0.4306	0.4722	0.4167	0.3472	0.4861	0.4583	0.5000
	7	0.4028	0.3889	0.3889	0.3333	0.4306	0.3750	0.4722
	9	0.3889	0.3611	0.3750	0.3611	0.4306	0.4306	0.4306
	11	0.4444	0.4167	0.3889	0.3611	0.4167	0.3611	0.3889
	13	0.4167	0.4028	0.4028	0.3333	0.4444	0.3333	0.3750
Error at individual level	1	0.5000	0.5556	0.5556	0.5000	0.6111	0.5556	0.5556
	3	0.3333	0.2778	0.3333	0.3333	0.4444	0.3889	0.3889
	5	0.3333	0.3333	0.2778	0.2778	0.5556	0.3333	0.4444
	7	0.4444	0.2778	0.2778	0.2778	0.4444	0.3333	0.3889
	9	0.3333	0.3889	0.2778	0.3333	0.4444	0.3889	0.3889
	11	0.3333	0.3889	0.3333	0.2778	0.3889	0.3333	0.4444
	13	0.2778	0.3333	0.3889	0.3333	0.3889	0.3333	0.4444

7 markers

	k	ankle, knee, toe, troch, shoulder, wrist, elbow
Error at step level	1	0.5000
	3	0.4167
	5	0.4306
	7	0.4028
	9	0.3889
	11	0.3750
	13	0.3889
Error at individual level	1	0.6111
	3	0.3333
	5	0.3333
	7	0.2778
	9	0.2778
	11	0.3333
	13	0.3333

Appendix 5
Use Case 3: Actual formation reports & Reference formations

Actual formation reports (FIFA, 2018)



2018 FIFA World Cup Russia™

Actual Formation

Round of 16

Belgium - Japan 3 : 2 (0 : 0)

54 02 JUL 2018 21:00 Rostov-On-Don / Rostov Arena / RUS

65' In 8 FELLAINI / Out 14 MERTENS
65' In 22 CHADLI / Out 11 CARRASCO

81' In 16 YAMAGUCHI / Out 7 SHIBASAKI
81' In 4 HONDA / Out 8 HARAGUCHI

FIFA WORLD CUP RUSSIA 2018

Total time played 96'18"

56% - 34'45" Ball possession 44% - 27'09"

Actual formation

Direction of play --> <-- Direction of play

COURTOIS	1
ALDERWEIRELD	2
KOMPANY	4
VERTONGHEN	5
WITSEL	6
DE BRUYNE	7
R. LUKAKU	9
E. HAZARD	10
CARRASCO	11
MERTENS	14
MEUNIER	15
FELLAINI	8
CHADLI	22

1	KAWASHIMA
3	SHOJI
5	NAGATOMO
7	SHIBASAKI
8	HARAGUCHI
10	KAGAWA
14	INUI
15	OSAKO
17	HASEBE
19	H.SAKAI
22	YOSHIDA
4	HONDA
16	YAMAGUCHI

Did not play the entire match Substitution Sent off Actual formation: average player positions

FIFA does not vouch for the accuracy of technical statistics compiled in association with its competition results system partner. The statistics will be confirmed at a later stage by the FIFA Technical Study Group.

MON 02 JUL 2018 22:36 CET / 23:36 Local time - Version 1 Page 1 / 1

2018 FIFA World Cup Russia™

Actual Formation

Quarter-final

Brazil - Belgium 1 : 2 (0 : 2)

 # 58 06 JUL 2018 21:00 Kazan / Kazan Arena / RUS

46' In 20 ROBERTO FIRMINO / Out 19 WILLIAN
58' In 7 DOUGLAS COSTA / Out 9 GABRIEL JESUS
73' In 8 RENATO AUGUSTO / Out 15 PAULINHO

83' In 3 VERMAELEN / Out 22 CHADLI
87' In 17 TIELEMANS / Out 9 R. LUKAKU

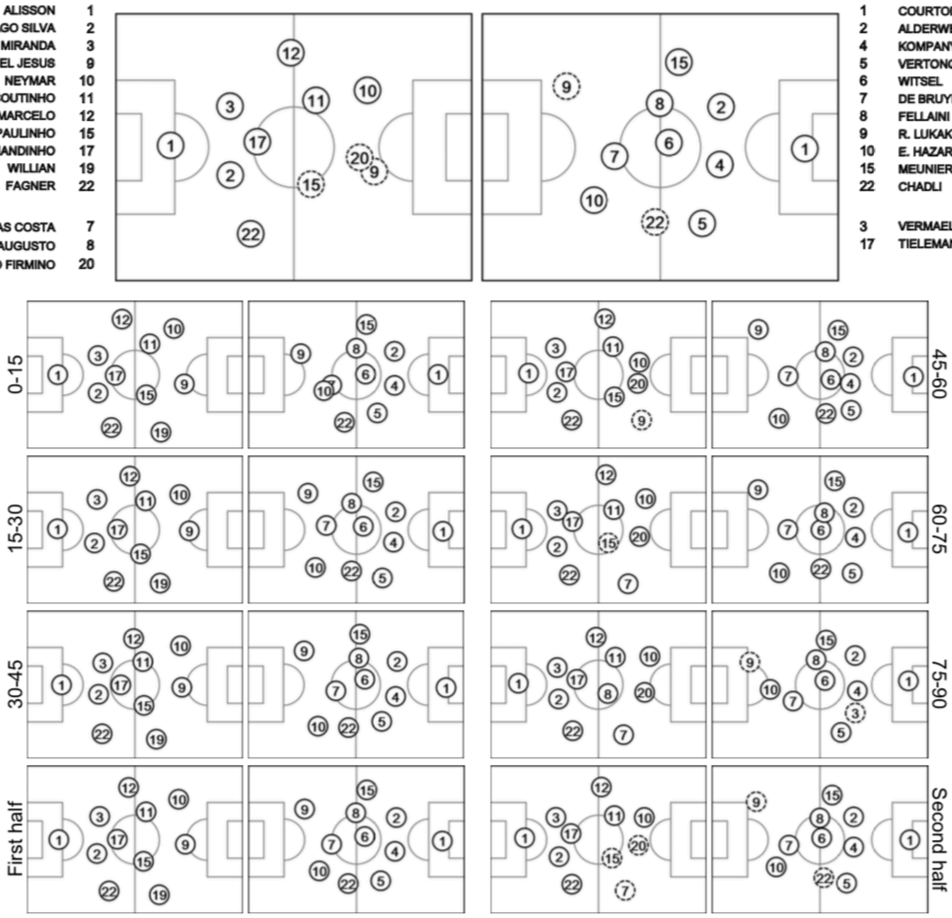
Total time played 96'12"

57% - 35'18" Ball possession 43% - 26'11"

Actual formation

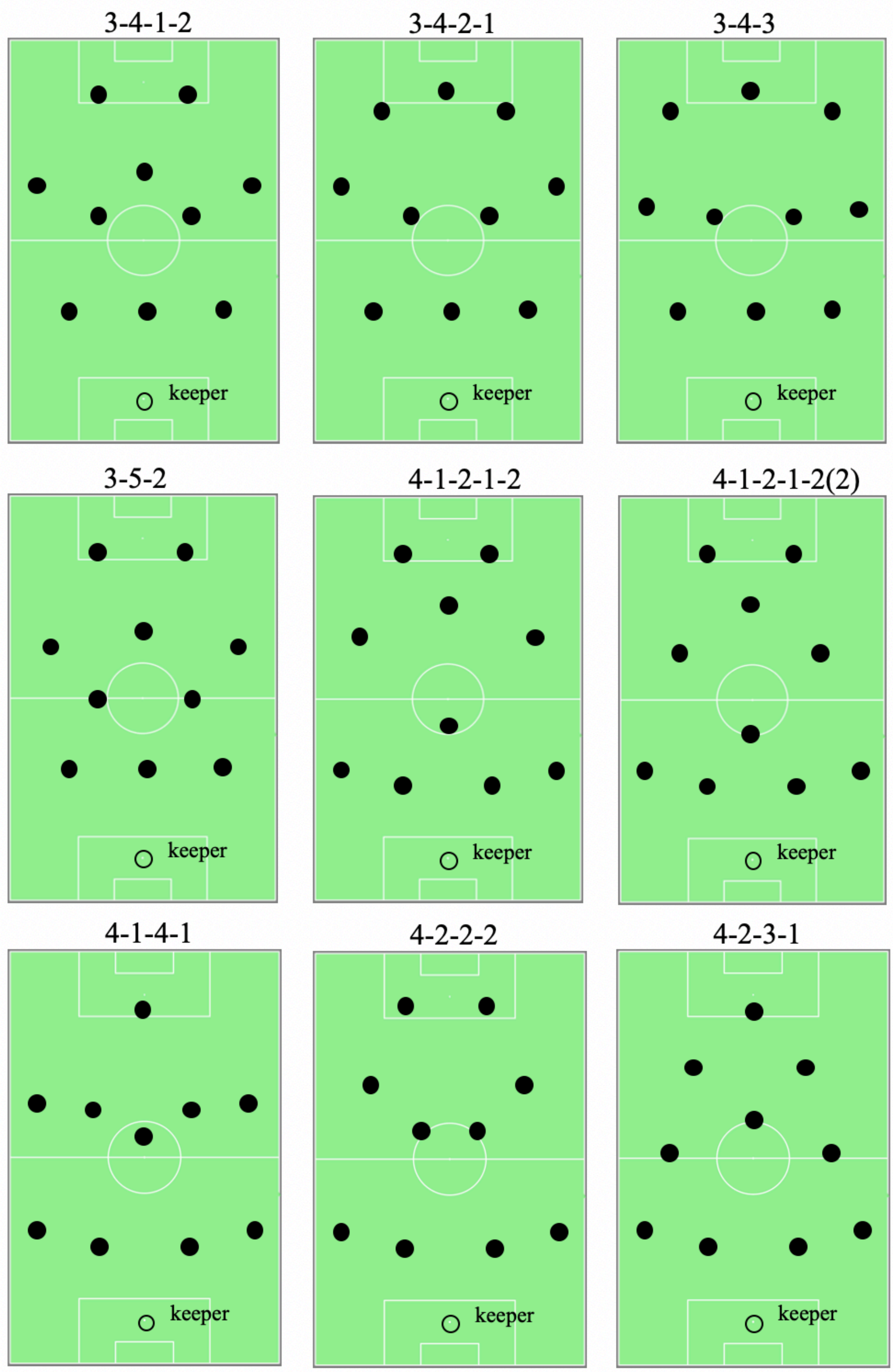
Direction of play --> <-- Direction of play

ALISSON	1	1	COURTOIS
THIAGO SILVA	2	2	ALDERWEIRELD
MIRANDA	3	4	KOMPANY
GABRIEL JESUS	9	5	VERTONGHEN
NEYMAR	10	6	WITSEL
PHILIPPE COUTINHO	11	7	DE BRUYNE
MARCELO	12	8	FELLAINI
PAULINHO	15	9	R. LUKAKU
FERNANDINHO	17	10	E. HAZARD
WILLIAN	19	15	MEUNIER
FAGNER	22	22	CHADLI
DOUGLAS COSTA	7	3	VERMAELEN
RENATO AUGUSTO	8	17	TIELEMANS
ROBERTO FIRMINO	20		

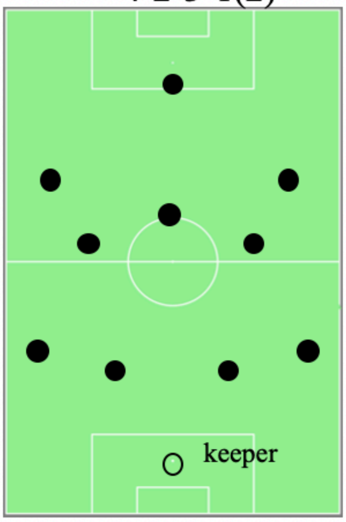


Did not play the entire match Substitution (R) Sent off Actual formation: average player positions

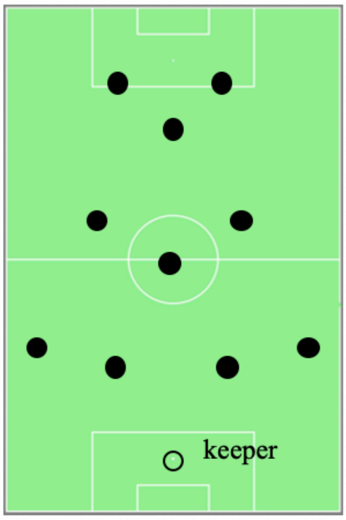
FIFA does not vouch for the accuracy of technical statistics compiled in association with its competition results system partner. The statistics will be confirmed at a later stage by the FIFA Technical Study Group.


3-4-1-2
keeper
3-4-2-1
keeper
3-4-3
keeper
3-5-2
keeper
4-1-2-1-2
keeper
4-1-2-1-2(2)
keeper
4-1-4-1
keeper
4-2-2-2
keeper
4-2-3-1
keeper

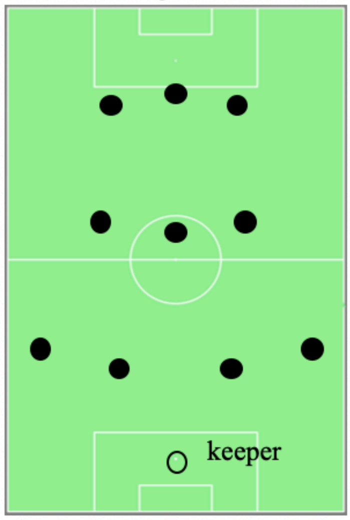
4-2-3-1(2)



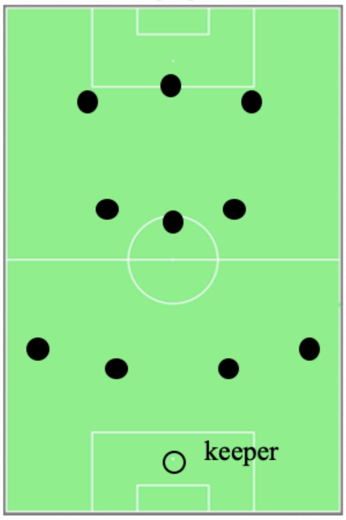
4-3-1-2



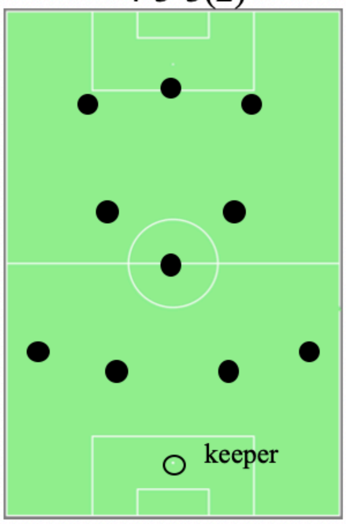
4-3-2-1



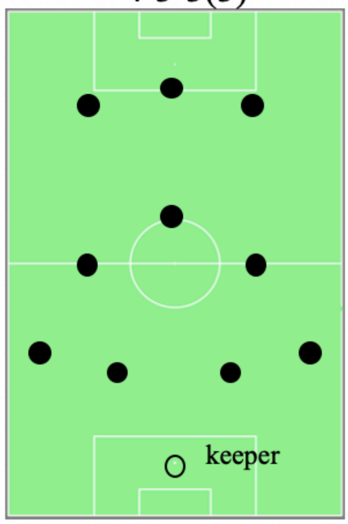
4-3-3



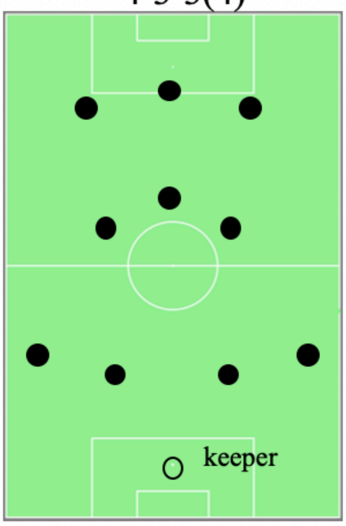
4-3-3(2)



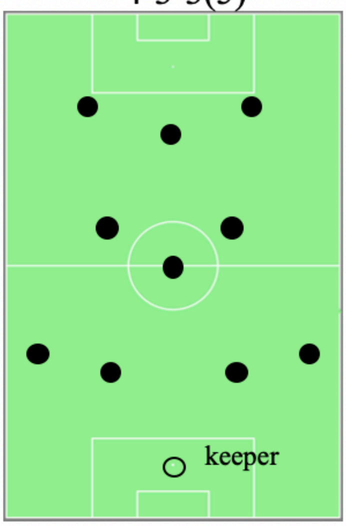
4-3-3(3)



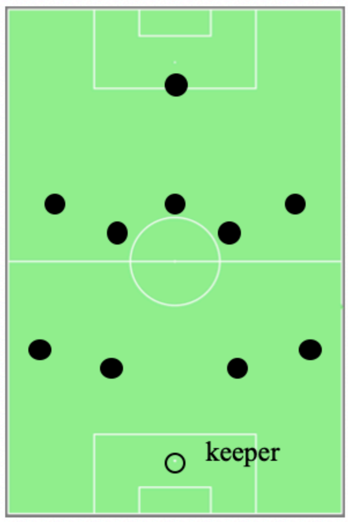
4-3-3(4)

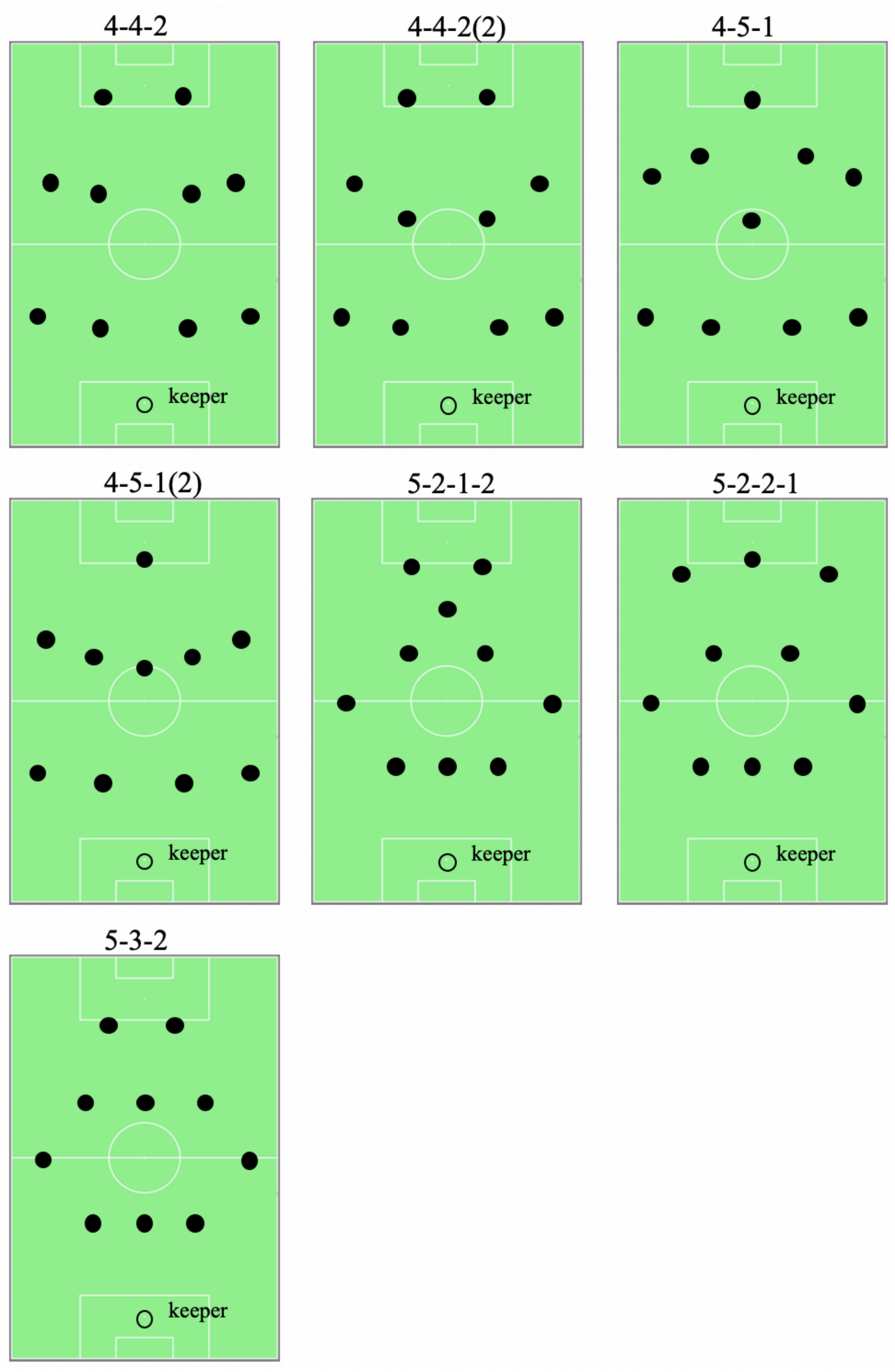


4-3-3(5)



4-4-1-1




4-4-2
keeper
4-4-2(2)
keeper
4-5-1
keeper
4-5-1(2)
keeper
5-2-1-2
keeper
5-2-2-1
keeper
5-3-2
keeper

Dutch Summary – Nederlandse Samenvatting

Het domein van sportanalyses kende een sterke groei in het laatste decennium. In het subdomein van sport performantie-analyse werden allerhande methoden geïntroduceerd, gaande van psychologische en sociologische analyses in teamsporten tot individuele statistieken over de ademhaling, het hartritme of zelfs de opname van nutriënten bij individuele atleten. Ondanks de verscheidenheid aan sporten met elk hun eigen karakteristieken hebben heel wat sporten een belangrijke gemeenschappelijke component: ruimtelijke verplaatsing door middel van wandelen of lopen. Het onderzoek dat in dit doctoraat gepresenteerd wordt situeert zich binnen het domein van de sportanalyse en heeft volgend doel:

De ontwikkeling van een methodologie voor de (geautomatiseerde) herkenning van sportpatronen door middel van de analyse van complexe ruimtelijke interacties tussen spelers met behulp van de van de Kwalitatieve Traject Calculus.

De Kwalitatieve Traject Calculus (*Qualitative Trajectory Calculus*; QTC) is een calculus die ontstond binnen de geografie en die relatieve bewegingen tussen bewegende puntobjecten beschrijft ten opzichte van elkaar door middel van kwalitatieve symbolen. De beoogde methode zou trainers, coaches en sport performantie-analysten de mogelijkheid kunnen geven om de ruimtelijke bewegingspatronen van een team te beschrijven en analyseren, om zo inzichten te verkrijgen in vaak gespeelde spelpatronen en de tactiek van een team. Op basis van het onderzoeksdoel worden volgende onderzoeksvragen geformuleerd:

R1 **Kan de Kwalitatieve Traject Calculus gebruikt worden om (op automatische wijze) frequent gespeelde bewegingspatronen te detecteren? (team gedrag).**

R2 **Kan de Kwalitatieve Traject Calculus gebruikt worden om (op automatische wijze) tactische patronen te herkennen in sport data? (team tactiek)**

R3 **Voor welke aspecten van sportanalyse kan de QTC-methodologie nuttig zijn, naast het analyseren van team gedrag en team tactiek?**

Bewegingspatronen in sport kunnen geanalyseerd worden zowel met een fijne temporele resolutie als met een grovere temporele resolutie (bijvoorbeeld door het uitmiddelen van de posities van spelers over langere tijdsperiodes). Daarom wordt er een onderscheid gemaakt tussen een methode voor dynamische sportanalyse en een methode voor statische sportanalyse. De eerste methode is gericht op de analyse van het team gedrag, terwijl de tweede methode de team tactiek analyseert. Aangezien QTC een reeds bestaande methode is die geïntroduceerd werd voor de onderzoekstelling van dit doctoraat, kan het onderzoeksopzet beschreven worden als de afstelling en kalibratie van de methode voor sportanalyse. De drie onderzoeksvragen kunnen dus verfijnd worden door middel van volgende meer specifieke onderzoeksvragen:

Ra **Hoe kunnen de QTC-representaties van sport intervallen (team gedrag) en sportmomenten (team tactiek) vergeleken worden? Hoe kunnen de gelijkenissen tussen de verschillende momenten of intervallen in sport vergeleken worden en wanneer zijn deze gelijkend? Hoe kunnen de resultaten van de similariteitsanalyse gevalideerd worden?**

Rb **Voor welke sporten kan QTC gebruikt worden als een methode voor sportanalyses?**

Rc **Hoe kan QTC verbeterd worden als methode voor sportanalyses?**

- **Wat is de impact van het gebruik van de verschillende QTC-varianten?**
- **Wat is de impact van het gebruik van statische punten?**
- **Wat is de impact van het gebruik van aangepaste grenswaarden voor de toekenning van QTC-karakters?**
- **Wat is de impact van het gebruik van de positionele data van de bal bij balsporten?**

Rd **Kan QTC gebruikt worden voor een meer gedetailleerde bewegingsanalyse, door het beschrijven en vergelijken van delen van het lichaam van individuele atleten?**

Dit doctoraat heeft als doel om antwoorden te geven op bovenstaande vragen en bestaat uit tien hoofdstukken. HOOFDSTUK 1 van dit doctoraat geeft een inleiding tot het domein van sportanalyse, de onderzoeksvragen en het onderzoeksopzet. Het vervolg van de dissertatie wordt opgedeeld in drie afzonderlijke delen. DEEL I bespreekt het gebruikte *framework* en bestaat uit drie hoofdstukken. HOOFDSTUKKEN 2,3 en 4 bespreken de theoretische achtergrond van de gebruikte methoden, terwijl in HOOFDSTUK 5 de implementatie van de methode in Python besproken wordt. De concepten die in de hoofdstukken van dit deel worden geïntroduceerd, worden geïllustreerd door middel van voorbeelden uit driebanden. Driebanden is een variant van biljart en heeft het grote voordeel dat het een relatief eenvoudige sport is: er zijn slechts drie ballen die bewegen in een gesloten omgeving (op een biljarttafel). Het bewegingsgedrag van de ballen wordt hierdoor, in contrast met het grote belang van menselijke factoren bij andere sporten, enkel bepaald door de initiële stoot op de bal en vervolgens door de wetten van de fysica. Om deze reden is driebanden een ideale sport om de concepten te verduidelijken, vooraleer ze toegepast worden op meer complexe sporten. DEEL II bestaat uit drie hoofdstukken (HOOFDSTUKKEN 6,7 en 8) die elk een gebruikscase bevatten van een specifieke toepassing voor sportanalyse. Elk van deze hoofdstukken bevat een discussie over het gebruik van de methode voor de specifieke sport. DEEL III van deze dissertatie bevat twee hoofdstukken. HOOFDSTUK 9 bevat een algemene discussie die kritische antwoorden probeert te geven op de gestelde onderzoeksvragen. HOOFDSTUK 10 geeft de algemene conclusie van het onderzoek. Referentielijsten worden afzonderlijk gegeven op het eind van elk hoofdstuk. Hieronder wordt een meer gedetailleerde samenvatting per hoofdstuk gegeven.

DEEL I - FRAMEWORK

HOOFDSTUK 2 geeft een overzicht van bestaand onderzoek binnen het domein van spatio-temporeel redeneren. De Kwalitatieve Traject Calculus (QTC), die geïntroduceerd werd in 2004 door Van de Weghe wordt hier in detail beschreven. Uitdagingen in verband met de toepassing van QTC voor sportanalyse worden aan het eind van het hoofdstuk gepresenteerd.

HOOFDSTUK 3 bevat de aanpassingen en toevoegingen aan de QTC-methode om deze bruikbaar te maken voor dynamische sportanalyse. Dynamische sportanalyse heeft als doel om frequent gespeelde bewegingspatronen te detecteren (team gedrag, R1). Het hoofdstuk verduidelijkt vooraleerst hoe QTC-matrix sequenties gebruikt kunnen worden om de QTC-relaties tussen meer dan twee spelers en voor meerdere tijdsintervallen op te slaan. Hierna worden de uitdagingen bij de toepassing van QTC voor dynamische sportanalyse besproken, evenals de mogelijke oplossingen: het gebruik van statische punten, aangepaste grenswaarden voor de toewijzing van QTC-karakters, het gebruik van verschillende QTC-varianten en permutaties van de QTC-matrices. De impact van deze aanpassingen wordt geïllustreerd door middel van eenvoudige drieband voorbeelden. Vervolgens wordt een nieuwe methode voor de berekening van de afstand (dissimilariteit) tussen de QTC-representaties van verschillende bewegingspatronen geïntroduceerd. Deze methode past de Levenshtein techniek toe op matrixsequenties van verschillende lengte en laat toe om bewegingspatronen met verschillende lengtes en/of met tempoverschillen te vergelijken. Op het einde van het hoofdstuk wordt de methode verduidelijkt door een uitgebreider drieband experiment, waarbij een reeks openingsstoten vergeleken worden.

HOOFDSTUK 4 bespreekt het gebruik van QTC als methode voor statische sportanalyse. Statische sportanalyse heeft als doel om tactische patronen te herkennen in sport op basis van de (uitgemiddelde) posities van spelers (team tactiek, R2). Initieel worden de aanpassingen en toevoegingen aan de QTC-methode verduidelijkt, die resulteren in een nieuwe variant (QTC_S). Hierna wordt de manier besproken om de afstanden tussen verschillende QTC_S-matrices te berekenen. Vervolgens worden de uitdagingen doorgenomen die gepaard gaan met de toepassing van QTC voor statische sportanalyse, samen met de mogelijke oplossingen. Deze bespreking is erg summier, gezien de gelijkenissen met de uitdagingen voor dynamische sportanalyse die uitgebreid aan bod kwamen in hoofdstuk 4. De oplossingen worden geïllustreerd met voorbeelden gebaseerd op driebanden.

HOOFDSTUK 5 geeft een overzicht van de implementatie van beide methoden. De twee afzonderlijke Python programma's (één voor dynamische sportanalyse en één voor statische sportanalyse) worden verduidelijkt door middel van gegeneraliseerde stroomdiagrammen. Vervolgens worden de verschillende programmatorische opties voor de implementatie van QTC besproken, samen met hun implicaties voor analyses van reële sportsituaties met QTC.

DEEL II – GEBRUIKSCASES

Het tweede deel van deze dissertatie bestaat uit drie gebruikscases die gepresenteerd worden in afzonderlijke hoofdstukken. Ze hebben als doel om de methoden voor sportanalyses te testen voor specifieke sporten en gebruikscases. Elke gebruikscase is ontworpen om informatie te leveren voor het beantwoorden van de onderzoeksvragen. Een evaluatie van hun waarde voor het beantwoorden van deze vragen wordt gegeven in deel III.

HOOFDSTUK 6 bevat een gebruikscase waarbij de methode voor dynamische sportanalyse wordt toegepast op voetbal voor de herkenning van bewegingspatronen van voetbalspelers (team gedrag). Het hoofdstuk geeft vooreerst een overzicht van de stand van zaken van onderzoek naar de herkenning van bewegingspatronen in balsporten in het algemeen, en meer specifiek in voetbal. Vervolgens worden de dataset en de setup van de gebruikscase toegelicht, waarna de resultaten beschreven worden. Om de significantie van deze resultaten aan te tonen bevat dit hoofdstuk een statistische validatie van de resultaten, gebaseerd op een voetbal expertenpanel.

HOOFDSTUK 7 bevat een gebruikscase waarin onderzocht wordt of de methode voor dynamische sportanalyse gebruikt kan worden voor een meer gedetailleerde analyse van de bewegingen van het menselijke lichaam. Hiervoor worden de wandelpatronen van kinderen met en kinderen zonder het *Development Coordination Disorder* (DCD) onderzocht. Het hoofdstuk begint met een literatuuroverzicht van relevant onderzoek naar het menselijke wandel- en loopgedrag. Daarna worden de dataset en de setup van de analyse verduidelijkt. Vervolgens wordt de methodologie die gebruikt wordt voor de statistische validatie van de resultaten geïntroduceerd. Hierna worden de resultaten en de validatie besproken. Het hoofdstuk wordt beëindigd met een discussie over de resultaten voor het domein van de bewegingsanalyse van wandel- en looppatronen.

HOOFDSTUK 8 bevat een gebruikscase waarin onderzocht wordt of de methode voor statische sportanalyse gebruikt kan worden voor een tactische analyse van voetbal, door het bestuderen van de teamformaties van de spelers tijdens een match. Het hoofdstuk begint met een literatuuroverzicht van analyse van teamformaties van balsporten in het algemeen en van voetbal meer specifiek. De gebruikscase bestaat uit een reeks van vrij eenvoudige (voetbal) voorbeelden en uit één meer diepgaand voetbalexperiment. Volgend op het literatuuroverzicht wordt een overzicht van de gebruikte datasets gegeven en wordt de setup van de analyses besproken. Vervolgens worden de voorbeelden van teamformatie analyse in voetbal gegeven, gevolgd door het experiment. Dit experiment bevat een validatie gebaseerd op de matchverslagen die gepubliceerd werden in populaire media. Het hoofdstuk eindigt met een discussie van de waarde van de methode voor statische sportanalyse voor de analyse van teamformaties in voetbal.

DEEL III – ALGEMENE DISCUSSIE EN CONCLUSIE

HOOFDSTUK 9 geeft een overzicht van de antwoorden op de onderzoeksvragen, gebaseerd op de resultaten van de gebruikscases van DEEL II. Naast een overzicht van de antwoorden worden de resultaten van het onderzoek kritisch benaderd in een discussie. Dit hoofdstuk bevat echter geen discussie van de resultaten van de gebruikscases voor de specifieke toepassingen (sporten), aangezien deze reeds in de desbetreffende hoofdstukken opgenomen werden. Tot slot geeft dit hoofdstuk nog enkele suggesties voor mogelijk onderzoek in de toekomst.

HOOFDSTUK 10 is een vrij kort hoofdstuk dat de nogmaals de belangrijkste resultaten van het onderzoek overloopt.

Curriculum Vitae

Jasper Beernaerts (°1991) was born in Ghent on the 2nd of October 1991. In 2009, he finished his high school education (Latin-Sciences) at Atheneum Voskenslaan in Ghent. Later that year, he started his education in Geography at Ghent University. Jasper eventually graduated as Master of Science in Geography (Summa Cum Laude) in June 2014. From October 2014 until September 2015 he worked as a scientific researcher at the Department of Geography of Ghent University on a project titled ‘Automated analysis of sports team movement tactics: development of a demonstrator and initiation of a licence track’. During that year, he applied for a PhD fellowship at the Research Foundation Flanders (FWO). After receiving a grant from the FWO, Jasper started his PhD research at the Department of Geography, CartoGIS research unit in October 2015.

Given the multidisciplinary character of Jasper’s research, the results of his PhD were published in (or are submitted to) diverse leading international journals in the field of Geography, Biomechanics and Biomechanical Engineering, Expert and Intelligent Systems and Sports Sciences. Besides articles in journals, Jasper received the best student paper award for his presentation on the paper titled ‘Analysing team formations in football with the static qualitative trajectory calculus’ at the 6th International Conference on Sport Sciences Research and Technology Support (icSPORTS 2018) held in Seville, Spain.