Using FIFA Soccer video game data for soccer analytics

Leonardo Cotta
Computer Science
Universidade Federal De
Minas Gerais, Brazil
leonardo.cotta@dcc.ufmg.br

Pedro O.S. Vaz de Melo
Computer Science
Universidade Federal De
Minas Gerais, Brazil
olmo@dcc.ufmg.br

Fabrício Benevenuto
Computer Science
Universidade Federal De
Minas Gerais, Brazil
fabricio@dcc.ufmg.br

Antonio A.F. Loureiro
Computer Science
Universidade Federal De
Minas Gerais, Brazil
loureiro@dcc.ufmg.br

ABSTRACT

Soccer has become the most popular sport in the world over the last century, but very little is known about its structure. One of the main reasons for that is the lack of a large-scale dataset. Despite the sport being constantly discussed and recorded worldwide, in every season there are several events that neither specialists nor fans can characterize or predict accurately. In this paper, we propose the use of a new dataset, the FIFA Soccer video game dataset, to better understand soccer. We describe the available data, justify its use and discuss possible applications. Finally, we show the potential of the dataset by analyzing two recent widely discussed topics, the contrast between the Brazilian and German National teams in 2014 and FC Barcelona's distinguished style in the 2012/13 season.

CCS Concepts

•Information systems → Crowdsourcing; •Human-centered computing → Empirical studies in collaborative and social computing;

1. INTRODUCTION

Soccer is the most popular sport in the world in both number of players and spectators [4]. In 2009, the market of global sports events was worth $64 billion, while soccer had 43% of it [2]. An important factor that contributed to its popularity is its unpredictability [5]. Unlike other sports such as basketball and baseball, the number of successful underdogs is considerably large every season in soccer leagues. In a context where underdogs are common, the rise of indisputable champions, such as Barcelona FC in seasons 2011/12/13 and the German National Team in 2014, also catches the attention of specialists and fans.

Despite of soccer's popularity, the prediction and characterization of matches and leagues outcomes are challenging open matters. The main reason for that is the lack of a large, unbiased dataset. Most of the knowledge generated in the area comes from former players or journalists bearing historical statistics about players and match results. The statistics used by soccer analysts have not shown much effectiveness in finding patterns among soccer teams and leagues, especially in large scale [3]. Usually, historical statistics characterize a specific aspect of the team or a player during the match, such as number of finishes or turnovers. As noted by NBA analyst Henry Abbot [1], those numbers usually correspond to actions that occur in a small fraction of time. A soccer match has at least 90 minutes and a player is active during all his stay in the field. Although on average a player runs from 10 to 15 km per match, he only possess the ball from 60 to 90 seconds [13].

To overcome these problems, data generated by sensors have been used to study soccer. Such data include videos from cameras to all kinds of physical measurements [8, 12] and human observations [7, 11]. The problem is that gathering such rich information about players may be very costly, making sensed data to be limited to teams with high purchasing power and unrestricted technology access. In addition, even if such approach was available to all teams, sensed data is always subject to physical interference. Therefore, the use of sensed data to analyze soccer may be unfeasible for worldwide large-scale studies.

In this paper we propose the use of video game data, more precisely from the FIFA Soccer series, from EA Sports[1], to better understand soccer. As aforementioned, it is very difficult to accurately characterize players. Nevertheless, if we have a precise characterization of them, we can better understand teams, matches, leagues and the sport as a whole. Fortunately, since 1995 the FIFA Soccer games provide an extensive and coherent scout of players worldwide. Each version of the game characterizes players with at least 20 different attributes and considers all the major soccer leagues in the world. For each attribute, we have an integer from 0 to 100 that measures how good a player is at that attribute. Examples of attributes are: *dribbling*, *aggression*, *vision*, *marking* and *ball control*. Observe that seems to be unfeasible to accurately characterize players in these attributes automatically. Thus, all of those are gathered and curated by payed workers of the company whose job is to bring the gameplay closer to reality as possible, hence preserving coherence and representativeness across the dataset.

Here we take the first steps to demonstrate that the FIFA Soccer dataset has the potential to enrich the field of soccer analytics. Moreover, to the best of our knowledge, this is the first work that proposes the use of video game data to analyze sports. In order to demonstrate the potential our proposed dataset, we analyzed two

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

[1] www.easports.com/fifa

recent and popular topics discussed by both sports media and fans. First, we studied the evolution of German and Brazilian elite soccer players throughout the years preceding the shocking semi-final match in 2014 FIFA World Cup between the two nations. Then, we analyzed the advent of a new and unique style in soccer, the so called *tiki-taka* in FC Barcelona's seasons of 2011/12/13.

2. DATASET

2.1 Dataset Description

The presented dataset relies on the principle of humans scouting players. In this work, we leverage the fact that EA Sports and other video game companies have been doing the scouting work for us over the last years. These companies have to model the players to be as realistic as possible in the game. This means quantifying their attributes, such as finishing and marking, and also mimicking their actions in the game by reality.

The FIFA Soccer series has over 500 licensed teams since the 2007 version, the leading game in this aspect. The company hires scouts all over the world to rate players' attributes as realistically as possible. Dubas-Fisher [6] shows how the ratings of FIFA Soccer attributes differ from conventional stats, therefore, distinguishing our dataset.

FIFA Soccer has a new game version for each new season. Each game version is released in the beginning of each season, around August or September. For instance, the game FIFA Soccer 07 refers to season 2006/07 in soccer. For each game version, each player has rating values associated with each attribute. As online gaming became a trend in console platforms, players' attributes ratings began to be updated during the seasons. The oldest version used in this paper, FIFA 07, has only two updates for each player, while the more recent versions are updated on a weekly basis.

Another evolutionary aspect of the game series is the number of attributes represented in each version. Over the last years, EA Sports has increased the number of attributes due to gameplay improvements. For instance, in FIFA Soccer 07 there are only 25 attributes, while in FIFA Soccer 16 there are 34. Nevertheless, we stress that although the attributes have increased in number, no attribute has been removed or merged. Therefore, we can perform comparative studies between different seasons by only considering the attributes from the oldest one. A full description of what skills each attribute represents can be found at EA's official website[2].

2.2 Dataset Collection

With the advent of the Internet, video games moved away from being played locally among friends in the living room or through LAN parties and started to connect thousands of players from all over the world in real time. Because of that, large communities of video game players emerged and worldwide competitions attract the attention of millions. As a consequence, communities of players are organizing and gathering all types of information about the games they are playing. The FIFA Soccer community is highly interested in knowing which are the best players to put in their teams. To do that, it is necessary to know each players' attributes and how they change over time. Thus, a natural consequence of this collective need is the appearance of online systems that show this kind of information in almost real time.

For this work, we collected the dataset from the SOFIFA[3] website, which is one of these systems and very well known by the FIFA Soccer community. It keeps record of all the players attributes ratings since the FIFA 07 version (including updates). SOFIFA[3] has also been used for consulting about the game in the media[6]. We crawled SOFIFA[3] in October 2015 and collected all the available information about players until then. We confirmed the veracity of the information by checking 10 random players of each game version from FIFA 07 to FIFA 16. As all of their attributes matched the ones in the website, we opted to trust their database. As part of this work, if this paper is accepted the authors commit to make the dataset publicly available.

3. DATASET APPLICATION

In the following sections, we show how the proposed dataset can aid in the analysis of two popular and recent stories in sports. For both of them, we explore the *overall* attribute, which is a summary of the players' attributes describing to the user how good a player is in one dimension. EA Sports does not explain how it is calculated, but we believe it is a weighted average of all the players attributes.

3.1 The Rise of Germany and Fall of Brazil

The match between Brazil and Germany in the 2014 FIFA World Cup was a shocking event, perpetuating discussions among fans and analysts. The match had the biggest winning margin in a World Cup semi-final or final. Besides that record, both Brazil and Germany are references in soccer. Brazil has five world cup titles and Germany, later on the competition, won its fourth title. During the game, by half-time, the score was already 5-0 and the German players seemed to be easily overpowering the Brazilian players. It is unclear until now exactly why Germany overpowered the famous Brazilian National Team so easily. It was a landmark in soccer history, evidencing the rise of German soccer and the fall of the Brazilian.

To better understand how these teams evolved over the years until the day of that historic match, we look at the 8 years preceding the 2014 FIFA World Cup. How does the Brazilian and German players have been evolving at each attribute? Did the defensive players evolve differently from the forwards and midfielders? To answer those and other related questions, we first classify each player in one of the following three categories: Defender, Midfielder or Attacker (Forward). Then, for each season, we computed the mean *overall* attribute of each player considering all his updates. Because Brazil and Germany had different squad formations throughout the season, we characterized each squad in a season by their top 20 *overall* players in each category. Again, for each of these players, we computed their mean attributes ratings over all his updates in that season. In summary, we have for each season and for each squad the mean attributes ratings of the top 20 *overall* players in each of the three categories.

We computed the aforementioned procedure for each of the 8 considered seasons, 2007–2014. Afterwards, we performed a linear regression model to each attribute rating over time, as illustrated in Figure 1 for attribute *Finishing* in the forwards' category. By doing so, we computed the slope coefficient to check if the best players of that country are getting better or worse at each attribute over the years. Moreover, we check which squad arrived better at 2014 World Cup by looking at the players' average attributes in 2014. Overall, the results are summarized in Figures 2, 4 and 3. The horizontal axis corresponds to the Brazilian players' slopes, while the vertical axis corresponds to the German slopes. The equality line separates attributes that Germany improved more than Brazil, and vice-versa. Points over the equality line indicate that both squads evolved equally in that attribute. Points above (bellow) the line mean that German players evolved more (less) at that attribute than Brazilian players. Additionally, we colored each attribute corre-

[2] www.ea.com/uk/news/the-backpage-fifa-12-attributes-guide

[3] www.sofifa.com

sponding to which team arrived better at it in 2014. Green squares indicate attributes that Germany had a better score than Brazil in 2014. Red diamonds indicate attributes that Brazil had a better score than Germany in 2014.

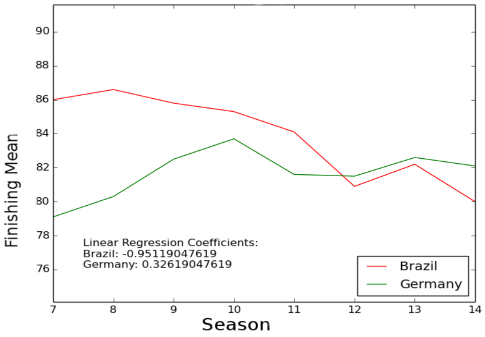

Figure 1: Top 10 forwards Finishing evolution for both Brazilian and German teams

Observe in Figure 2 that Brazilian defenders evolved more than Germans in most attributes, but mostly in non-defensive[4] ones, such as *Crossing* and *Dribbling*. In Figure 3, we note an apparent balance in the attributes' evolutions and values in 2014, but German players have evolved more in the offensive attributes, such as *Finishing* and *Shot Power*. The main German advantage is seen in Figure 4. German midfielders evolved more at most attributes, all of the offensive ones, and arrived better at them in 2014. These observations, made from our proposed dataset, suggest that the outcome of this particular World Cup match in 2014 had been shaped years before, as German players evolved more consistently and efficiently (i.e., in attributes that matter) than Brazilian ones. Of course this is not sufficient to accuse the shocking 7–1 result of this match, but it makes clear the disparity between these two teams at that time.

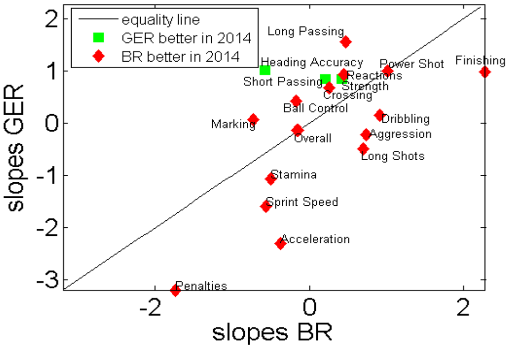

Figure 2: Top 10 defenders attributes coefficients

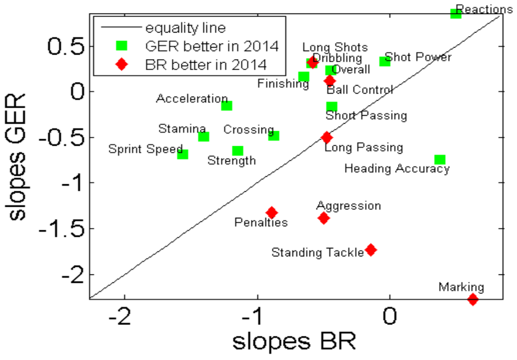

Figure 4: Top 10 midfielders attributes coefficients

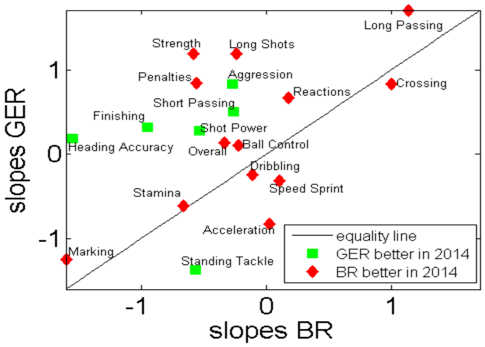

Figure 3: Top 10 forwards attributes coefficients

3.2 FC Barcelona's Tiki-Taka

One of the recent revolutions in soccer was made by Head Coach Pep Guardiola and his FC Barcelona squad. From 2008 to 2012, he proposed a new style in soccer, the so called *tiki-taka*. By preserving dominant ball possession and constant passing strategies, his team revolutionized and dominated the sport in these years.

Gyarmati et al. [7] showed how the passing patterns in Barcelona FC differed from the other teams in the Spanish league, evidencing a unique style. However, fans and specialists still wonder whether the unique style is related to a unique coach, a unique squad or the combination of both. Can any squad reproduce the *tiki-taka*? To approach that question, we analyze the same squads analyzed in [7] using the proposed FIFA Soccer dataset .

For the 2012/2013 Spanish league squads, we only considered players that at some point in the season were either a starter or a substitute. As the ball stays mainly with the midfielders during a game, who account for most of the passes, we only considered the squads' midfielders. Again, for each team, we computed the mean of each attribute rating for their midfielders. In order to consider the midfielders' characteristics profile, instead of considering their absolute mean values, we normalized each mean attribute rating by the mean *overall* of that midfielders' squad. Therefore, teams that naturally have lower or higher attributes ratings because, for instance, their economic power, would not differ in such way. Instead, this normalization procedure accounts for how much an attribute influences in the midfielders' *overall* attribute.

As [7], we performed a Principal Component Analysis[9] in our features for each team. Afterwards, we performed a k-means clustering[10] using the first 19 components that, in total, account for

[4] Attributes that do not contribute directly to a defender efficiency

more than 99% of the variance. Figure 5 shows the clusters we found from this procedure. First, observe that a cluster has only one team, FC Barcelona, the same phenomenon observed by [7]. Here we note that more than a unique style, FC Barcelona had unique midfielders. The aforementioned results can also indicate why Pep Guardiola has not been able to reproduce the success he had in FC Barcelona in FC Bayern München recently. Or even how the Spanish team apparently reproduced the same style in 2010 FIFA World Cup with an almost identical midfield. As part of a future work, the authors plan to investigate which specific features differentiate Barcelona's midfielders from other teams, such as Guardiola's Bayern München.

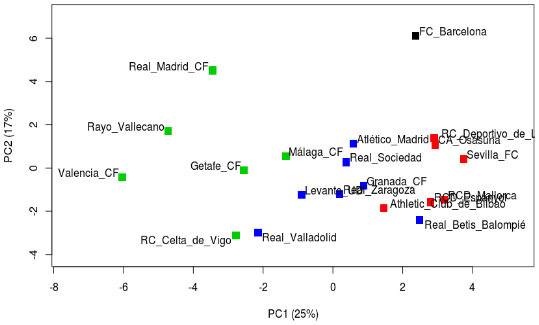

Figure 5: K-means clustering of the teams in the Spanish league. One of the four clusters contains only a single team, namely, FC Barcelona that has unique midifielders

4. CONCLUSION

In this paper, we showed the potential of online sports video games data to provide large-scale and unbiased studies. We proposed the FIFA Soccer series dataset, explained its structure and potential use. Furthermore, we analyzed two widely discussed topics to show how this dataset is able to give valuable insights concerning these topics. We believe this work can lead to other relevant studies, where sports video game series can be explored to generate knowledge in sport. While this work focused on soccer and FIFA Socccer video game series, other sports and video games systems can be used in the same sense.

5. REFERENCES

[1] H. Abbot. Bad use of statistics is killing anderson varejão. *True Hoop*, november 2007.

[2] ATKearny. The sports market. www.atkearney.com/en_GB/paper/-/asset_publisher/dVxv4Hz2h8bS/content/the-sports-market/10192, 2009.

[3] C. Bialik. Can statistics explain soccer? *The Wall Street Journal*, June 2008.

[4] E. Britannica. Football. www.global.britannica.com/sports/football-soccer, Aug. 2015.

[5] A. Creditor. Unpredictability running rampant in world cup's toughest groups. www.si.com/soccer/planet-futbol/2014/06/20/world-cup-unpredictability-costa-rica-usa-concacaf-italy-england-uruguay/.

[6] D. Dubas-Fisher. Fifa ratings vs real life: How does the video game measure up to actual premier league stats? www.mirror.co.uk/sport/football/news/fifa-ratings-vs-real-life-4027925.

[7] L. Gyarmati, H. Kwak, and P. Rodriguez. Searching for a unique style in soccer. *arXiv preprint arXiv:1409.0308*, 2014.

[8] P. Halvorsen, S. Sægrov, A. Mortensen, D. K. Kristensen, A. Eichhorn, M. Stenhaug, S. Dahl, H. K. Stensland, V. R. Gaddam, C. Griwodz, et al. Bagadus: an integrated system for arena sports analytics: a soccer case study. In *Proceedings of the 4th ACM Multimedia Systems Conference*, pages 48–59. ACM, 2013.

[9] I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[10] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[11] J. L. Pena and H. Touchette. A network theory analysis of football strategies. *arXiv preprint arXiv:1206.6904*, 2012.

[12] D. Schuldhaus, C. Zwick, H. Körger, E. Dorschky, R. Kirk, and B. M. Eskofier. Inertial sensor-based approach for shot/pass classification during a soccer match.

[13] R. Tucker. Physiology of football: profile of the game. http://sportsscientists.com/2010/06/physiology-of-football-profile-of-the-game/.