A new massive data analytics approach to the football

Marco De Nadai, Michele Linardi

ABSTRACT

This paper approaches the football with a similar way as Moneyball has made in US baseball. It wants to create a small framework of tools which permits the coaches to base their new strategies not only on their intuitions and judgments but also on comprensive (and maybe not human-eye-visible) statistics. Using the past researches in the football and sport fields, it will analyze a match recorded by DEBS with an innovative system of sensors which could be the future of football.

1. INTRODUCTION

In a modern (or post-modern) society where the hero-cult, self-improvements and the game are important, and where the multi-etnics/culturality are increasing, the sport influence is uncontested [19]. Football is the most popular sport in the world but, while the national and international competition increases, the investments need to diminish due to the high debts reached from all the major football clubs (only in the italian Serie A there are 1.6B euro of debts [8]]). It is clear that without new money and investments, it is not possible to follow the self-serving megalomania [23], which conducted these clubs to this point. They need to rely in alternative system.

This situation is very similar to what the Okland Athletics' general manager, Billy Beane faced in 2002: Okland Athletics (OA) was in an unfavourable financial situation after 2001 postseason. For this reason, he focused his attention on the teams analytical, evidence-based SABS [1] approach to assemble a competitive team which challenged teams such as the New York Yankees, who spent three times more than OA in payroll that same season. His approach, well described in the book *Moneyball, the art of winning an unfair game*, was based in tricky stats which contributed in the selection of player who could have leaded to the World series victory. Bill James condensed the SABS' reasoning on this quote: "sabermetrics attempts to answer objective questions about baseball, such as "which player on the Red Sox contributed the most to the team's offense?" or "How many home runs will Ken Griffey hit next year?" It cannot deal with the subjective judgments which are also important to the game, such as "Who is your favorite player?" or "That was a great game." [12].

After a utopic OA season, Moneyball has made an undeniable impact in the baseball major league, hence the most important teams have hired full-time SABS analysts and have started to follow its principles. Although football is related to his "stop and go" nature, in 2009 *Soccernomics* approached the European football as Moneyball did for U.S. baseball, influencing the selection of players in a team. However, the other reason to analyse players is linked to the exploitation of the factors, which contribute to optimal players' performances: examining these factors, coaches could understand camouflaged problems and train better the players. They could analyse a match to identify the major team's strengths, which can be further developed, or its major weaknesses, which suggest areas for improvement. Similarly, the same analysis applied to the opposite team, could make the strategy development easier to the coach [16]. The increased research activity in this field has been particular evident in soccer, where the importance of scientific research has became increasingly accepted [21].

The player monitoring during a competition were originally achieved using manual video-based motion analysis, but the perceived complexity and time consumption formed barriers to its adoption [14]. In the last decade, teams often used automatic motion analysis realized with system based on PITCHf/x [4] with many limits, errors, problems [5][10] which did not lead to meaningful results and consequently did not justify they high maintaining costs.

The recent 2013 MIT Sloan sport analytics conference emerged [4] the rise of spatial data, which are extremely important in every sports to explain trickily problems inside a team with a high accuracy. The use these wireless GPS systems to track players movements during a match, connected with high quality data visualization makes complex data analysis interesting, useful and understandable to GM and business people. The intense competitive schedules of elite soccer clubs require data to be available usually within 24-36 hours post-match [motion analysis citation], for this reason we decided to use the DEBS 2013 conference [6] challenge [7] dataset to analyse a football match and the players involved. We will use the recent data mining technologies in order to exploit different meaningful tools to the coaches.

2. RELATED WORKS

When Bill James and SABR-metrician were making their contributions to baseball, they worked in public data. Nowadays things are different: sport analytics is made primarily in-house, were number crunchers can work secretly in order to gain a competitive advantage. This peculiarity is true especially in football, where companies like Opta in England, Adidas in USA and many others (Sport/VU, Prozone etc) are contracted to gather in depth stats to the teams who do not distribute their dataset. Despite that, there are some exception like Manchester City FC, who made available to the world the MCFC Anaytics project: some statistical data on the 2011-12 English Premier League.

3. PROPOSED APPROACH

The importance of teams is widely accepted [15][22] and the right composition of teams determines their odds or success [13][24]. The difficulty relies in the understanding what and who is important in a team.
Football analysis is not easy as baseball; indeed the first is a unique sport, characterized by continuous, cooperative and interdependent actions, hybrid offensive and defensive roles and thousands of thousands of variables dependents from the teams strategy. This makes very difficult to exploit what is important in a match in order to achieve a victory and forces researchers to analyse many matches and tournaments in order to discover similar patterns.

The easier statistic adopted in the past has been the possession percentage; to cite Johan Cruijff: "As long as we have got the ball, they can't score". However, this metric is gradually losing ground as the recent researcher states [20]. In the researches so far, efficient attacks were analysed in the European Championships, World competitions/championships, high quality club competitions but these studies' outcome are contradictory. Indeed for example, if talk again about possession, Hook and Hughes (2001) found that successful teams utilised longer possession than unsuccessful teams in Euro 2000; instead in a similar study Stanhope (2001) found that time in possession of the ball was not indicative of success in the 1994 World Cup. Other studies have tried to provide a "formula" of winning but anyone came out with a definitive conclusion [16]. Nonetheless, the latest research in this field focuses in the network analysis to study the relations of players in the matches, as a system biology group explained citeduch2010quantifying. Hence, we concluded that a "final formula" does not exist and statistics needs to be helpful to coaches and not a substitute to them. For this reason we will analyse four different characteristics from the DEBS challenge: similar players to train, passages patterns, clustering of trajectories and apeed perfornance analysis.

3.1 The dataset

The purpose of this work is exploiting some useful hidden informations of a football match, using the data recorded by the sensor provided by Fraunhofer, used in the challenge promoted by the [6] conference. This dataset is composed by a set of positions of players and balls which records the activity of a soccer match.

The event schema of the sensor recording is following: `sid, ts, x, y, z, |v|, |a|, vx, vy, vz, ax, ay, az` where `sid` is a sensor id which produced the position event, `ts` is a timestamp in picoseconds, e.g.: 10753295594424116 (`x`, `y`, `z` describe the position of the sensor in mm (the origin is the middle of a full size football field) ; `|v|` (in m/s), `vx`, `vy`, `vz` describe direction by a vector with size of 10,000. `ax`, `ay`, `az` describe the absolute acceleration and its constituents in three dimensions (the acceleration in m/s is calculated similar to that of the velocity). The acceleration does not include gravity, i.e. `|a|` is zero when the ball is at a fixed position and not 9.81 m/s).

3.2 Similar players to train

Past research focuses [16] concluded that the variables that better differentiate winning, losing and drawing teams in a global way were the following: total shots, shots on goal, crosses, crosses against, ball possession and venue (fixed in our dataset). Despite that it seems clear that the ability to retain possession of the ball is linked to success, we will not consider the possession time due to the problems described before; we will only consider the players effectiveness. Moreover it is also proved that the second half of the match is significantly more efficient and the efficient attacks are initiated mostly by interception of opponents passes [20], so we counted the tackles, weighting them considering the first/second half of the match. All the weights are based on the multivariate test of [16], and it conduct to a multidimensional value which summarizes the player quality. Afterthat we will group the players into similar clusters.

Our Python code reads the whole dataset with the csv reader/writer module, which permits a good organization and balance between the memory and I/O used. In the code we start excluding every sensors which are out from the field, afterthat we observe either the ball and the players. Some game considerations/euristics are necessary:

- A ball is hit if it reaches the minimal acceleration of 55 m/s
- A player possess the ball if the distance from one of his foot (transmitter) is less than 1 meter and he keeps it in a time between 0.2 seconds and 1 second.
- The ball is crossed if the z-coordinate is greater than 2 meters.
- There is a tackle if the team possession changes in a short time and between two players sensors within 2 meters of distance.

The code analyses some dimensions obtained by different different reasearches in the football field. Consequently the analyzed metrics are:

successful/unsuccessful passages important for the ability to retain possession of the ball, linked to the effectiness of attack/defense

unsuccessful passages on defensive third dangerous errors near the goal

successful/unsuccessful crosses important actions for an effective attack [16]

total shots and shots on goal important actions for an effective attack [16]

won/lost tackles important for the interceptions and the consequently construction of the actions, especially in the second half of the match [1]

Afterthat, the code will save the results into a transition SQLite database which can be used many times with different algorithms and parameters without lunching the analyzation again.

The second part of the work is addressed to give mean to the numbers, clustering the similar players into groups and attaining high intra-cluster similarity (documents within a cluster are similar) and low inter-cluster similarity (documents from different clusters are dissimilar).
A number of clustering algorithms with different characteristics has been reported in literature, hence we examinated a considerable number of them, including hierarchical and partition clustering algorithms. The first group seeks to build nested clusters by merging them successively. It is divided in two main categories: agglomerative and divide. They are very useful for our project because they use a distance matrix to compare the similar object in a set and they need only a termination condition. Partition-based clustering algorithms' aim is to partition the observations into k clusters in which each observation belongs to the cluster with nearest mean.

The hierarchy-clustering algorithm selected is Ward which is a variance-minimizing agglomerative clustering. The purpose of this algorithm is to unify groups such that the variation inside them does not increase too drastically. Specifically it joins the groups that give the smallest increase in:

$$\Delta(P,Q) = \frac{n_P n_Q}{n_P + n_Q} d^2(P,Q)$$

This procedure will make clusters as homogeneous as possible. This type of clustering is used in our work in order to visually plot a dendogram, which specify the similarity of the groups. Coaches could form groups or think about strategies visually.

The partition based clustering algorithm we will use is K-means [11], which requires the k number of clusters we want in output. It starts identifying k cluster centers and then it calculates the distances of observations from the k-centers and assigns each observation to the nearest center, forming a group. After that it calculates the new center of each group and start again to assign the observations to the nearest cluster center. The procedure stabilizes when none of the N objects changes the membership. The major weaknesses of this algorithm are not significant to our work:

- the number of clusters, which needs to given in advance could be the number of available coaches of a team. For example if one club has three coaches, it could train its players grouping them by similarity. This could make the workout very efficient.
- Its sensitivity to noisy data and outliers is not a problem because we need to train every teams player: no one could avoid the workout.

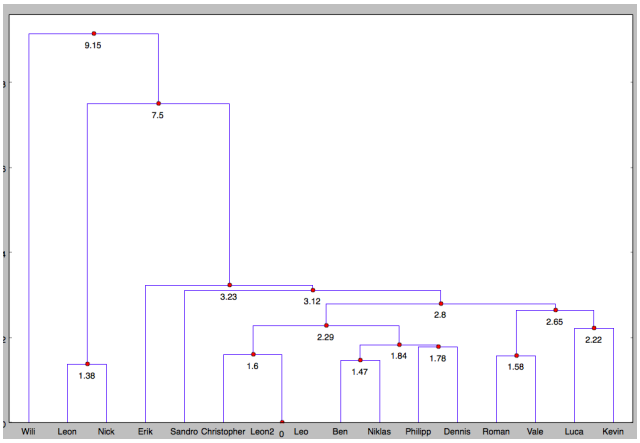


Figure 1: Dendogram of the hierarchical clustering

As mentioned above, the algorithm needs to locate the first k centers and it could be done randomly. We will use instead the k-means++ heuristic [2], which iteratively chooses the next seed with probability depending on how far apart it is from any of the currently selected seeds.
The result of k-means clustering is composed by the k-clusters of players and the mean of their weaknesses, useful to address the training.

3.3 Passages patterns

This work is based on the methodology created by a group of system biology, which studied the importance of players, and those passes in the Euro 2008 tournament [9]. In order to measure the player importance in a team, each player is represented in a graph with a node and the passages between the different players with an edge. The darker the node colour is, the more efficiency he reached. Similarly the thicker the edge is the more number of passages he did with the linked player. This result is strongly connected to the previous tool we developed and we refer to it for further explanations.
The output graph could be used by coaches to valuate the importance of players (and superstar-players) into their team. Similarly it can be use to build a strategy for an incoming match.

3.4 Clustering of trajectories

In order to make good strategies the soccer coaches analyze the archives of matches. Since the soccer is a very complicated game due to a large number of participants for each team and for the several types of existing roles, in our opinion, understand the dynamic of the movement of a single player could be meaningful for several problems. First of all discover the recurrent patterns of a player movements is useful to understand how the player usually act and this can reveal two several important things:

1. Discover if a player is compatible with a module or with the type of play that the coach wants to adopt.
2. Discover if two or more player have the same movements.

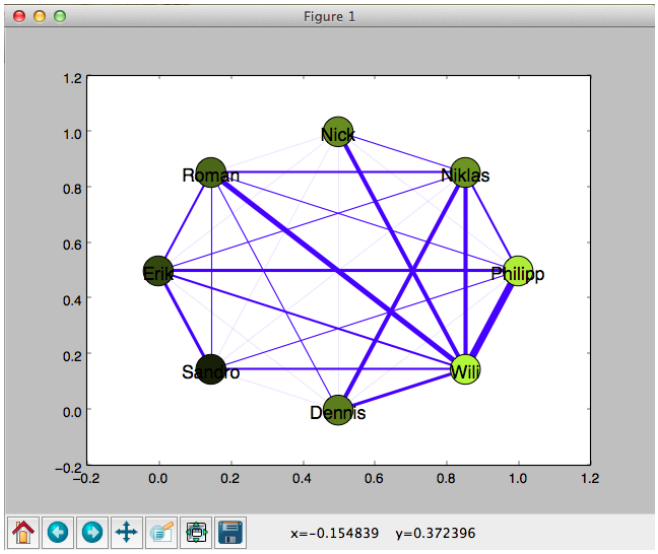


Figure 2: Passages patterns of team 1

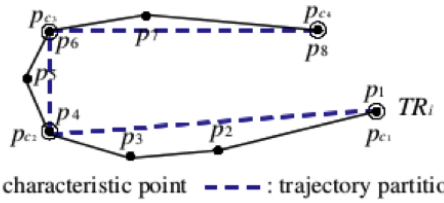
3. Understand the movement of a striker in order to apply the best defensive strategy, and of course the symmetric problem (understand how the defenders move to discover the portion of field where develop the action).

In order to formalize this problem we can state the follow main points:

1. Find the set of all significant trajectories for each player
2. Find the set of all significant trajectories for the ball
3. Cluster (group) all the similar trajectory of a single player or the similar trajectory of two or more player. In the same fashion, we will cluster the entire trajectory for the ball.

3.4.1 *Find the meaningful trajectory*

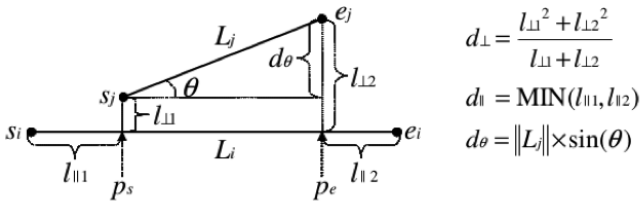
Now we can define the first point related to the discovery of significant trajectory of players and we will show how we extract them in real time from the sensor data. The first extraction from the data aims to store the trajectory, recording the single point of the player every second (we assume reasonable to log the movement of the player with this frequency). At the end of this step we have a sorted set of point P = (x1,...xn) that tell us the activity of the player in the key moments of a match.
Now we can shift toward the second main problem that is to find the similar trajectory using an intuitive and powerful framework called TRACLUS [17]. The optimal partitioning of a trajectory should possess two desirable properties: preciseness and conciseness. Preciseness means that the difference between a trajectory and a set of its trajectory partitions should be as small as possible, conciseness means that the number of trajectory partitions should be as small as possible. In order to find the best set of characteristic point, in other words, the set of point where the trajectory has a significant change of behaviors the TRACLUS partitioning algorithm aims to find the best trade off between conciseness and preciseness.



The TRACLUS framework uses the MDL (Minimum description length) principle to minimize the trade off between conciseness and preciseness. Using this principle we have two values $L(H)$ that measure the degree of conciseness, and $L(D|H)$ that of preciseness. H is the hypothesis and it corresponds to a specific set of trajectory partitions and D the data thus the whole trajectory. $L(H)$ is the length, in bits, of the description of the hypothesis and $L(D|H)$ is the length, in bits, of the description of the data when encoded with the help of the hypothesis. $L(H)$ is computed summing the length of all the segment formed by the characteristic point. $L(D|H)$ is computed summing the difference between each segment with the whole trajectory. The difference is computed with the help of three distances:

1. Perpendicular distance
2. Parallel distance
3. Angle distance

These distances are used for determining the trajectory similarity during the clustering process as well. To compute $L(D|H)$ are used only the perpendicular and angle distance, the parallel distance is discard because the sub-trajectories belong to the same trajectory.



At the end we keep the partition that minimize the summation: $L(H) + L(D|H)$. We find that exploiting this method of partitioning we obtain a valid extraction of the meaningful trajectories of a player. We will see the experimental results that give us an important feedback about the quality of information gathered in the context of football.
At the end of the partitioning as mentioned above we run the clustering algorithm that accept in input all the sub-trajectories partitioned that represent the interesting movement. As the major of the density-based cluster algorithm, also the TRACLUS suffers the problem of the parameter depending. Specifically there are two parameters to be set:

1. Epsilon-Neighborhood, that permits to discovery the belongingness of a sub-trajectory to a specific cluster.

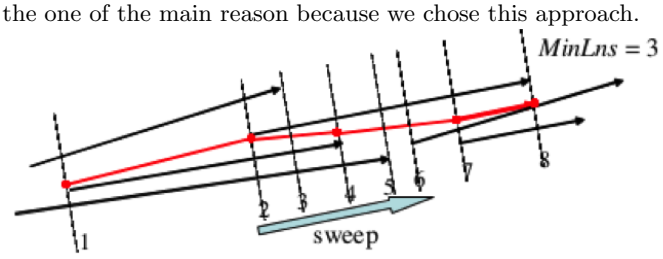
2. MinLns, that is a parameter telling us the minimum size of the clusters and the number of the original trajectories where the sub-trajectories have to come from.

Now we are going to describe how we solved the problem of the parameter discovery.

For the Epsilon-Neighborhood, we can exploit the method yield by the TRACLUS framework where it attempts to find the best value based on the entropy minimization. Of course, during the experimental evaluation, we tried to see how the quality of clustering is affected with the changing of the Epsilon parameter. Concerning the MinLns parameter we decided to set it based on how many player's trajectory we have. Making an example if we check the sub trajectory of one player is enough that the sub-trajectories come from only one trajectory that is the whole trajectory of the player. If we want to cluster the trajectories of more than 1 player we set the MinLns to the number of player so we are sure to find all the common sub-trajectories that come from all the player respecting the goal of the clustering that is to find similar common movements of the player. At the end to avoid small cluster we discard the cluster that contains only one trajectory when we cluster the trajectories of one player.

3.4.2 *The powerfulness of the representative trajectory*

After the clustering, what is interesting is the method of the representation of the clusters. The TRACLUS Framework yields a very expressive method to represent the cluster. More in details, it computes the average of the movements of the all sub-trajectories and it is very important to understand the data in the domain of football. This is for sure the one of the main reason because we chose this approach.



3.4.3 *Experimental evaluation of clustering*

Seeing the plot of the clustering we can get out with some reasonable and interesting result. The first result is the choice of the good parameter of Epsilon through a visual analysis. The TRACLUS estimation for the player's trajectories suggests a parameter of 20. However, we started to see how the quality of cluster was affected starting with a very low parameter (1) going ahead with 5, 10 and 15. We experimented that aspect on a striker after 4 minutes of match.

From a rapid visual investigation we can say that from 10 to 20 the quality of clustering does not change significantly and we feel free to assume that 10 for the football movement is a reasonable parameter where to around. Of course having only a dataset for one match we cannot confront with other matches. This suggests a cue for a future work that

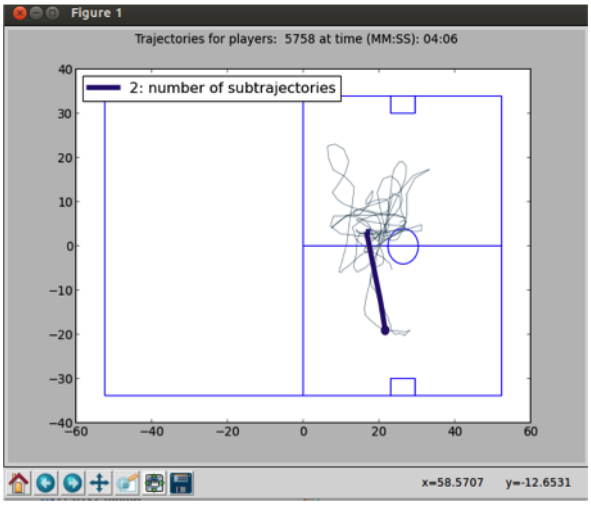


Figure 3: Striker after 4 minutes with a parameter of 1 (the cluster is represented by the bold line, instead with the thin line is represented the complete trajectory)

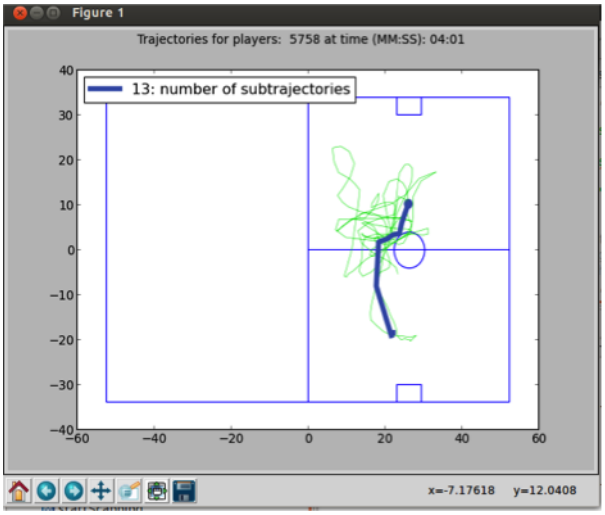


Figure 4: Striker after 4 minutes with a parameter of five

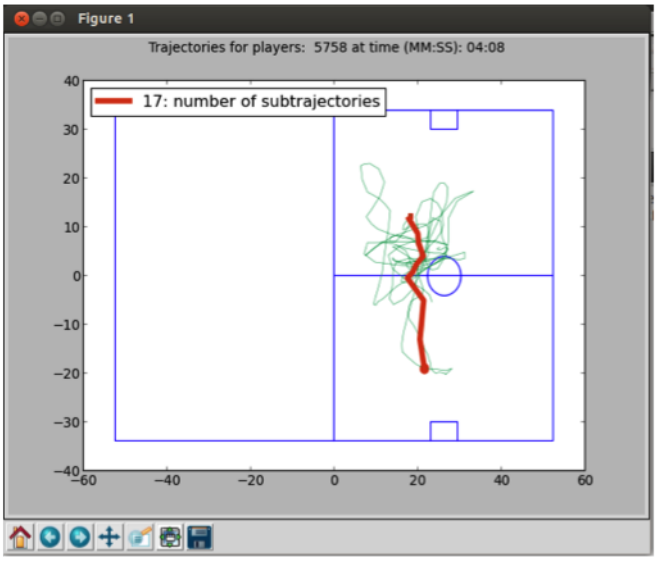


Figure 5: Striker after 4 minutes with a parameter of 10

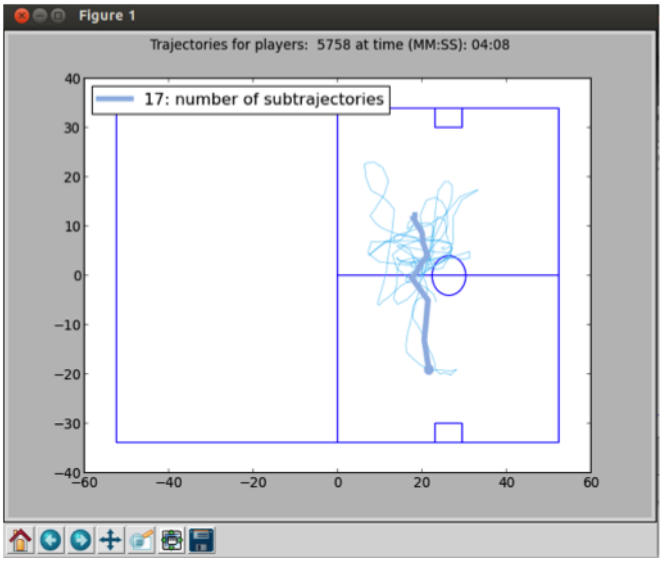


Figure 6: Striker after 4 minutes with a parameter of 15

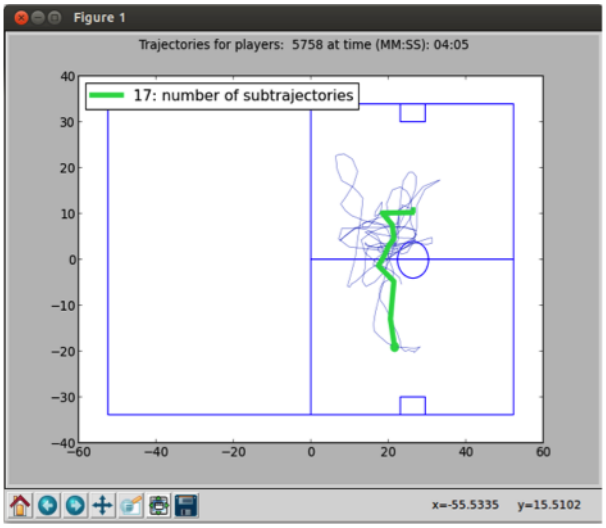
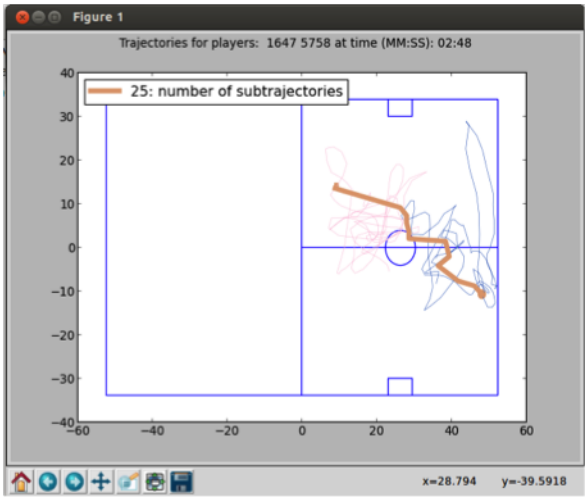
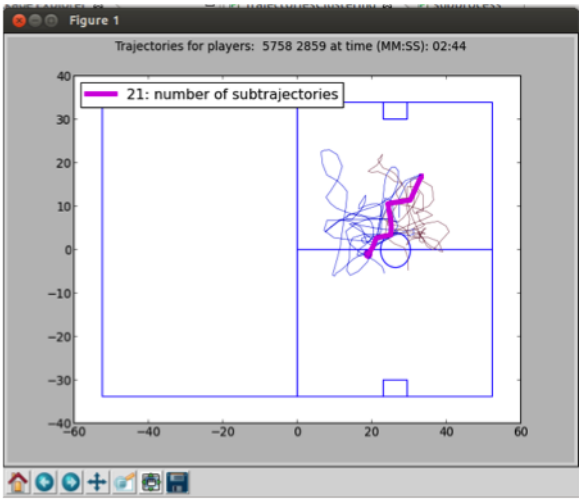


Figure 7: Striker after 4 minutes with a parameter of 20

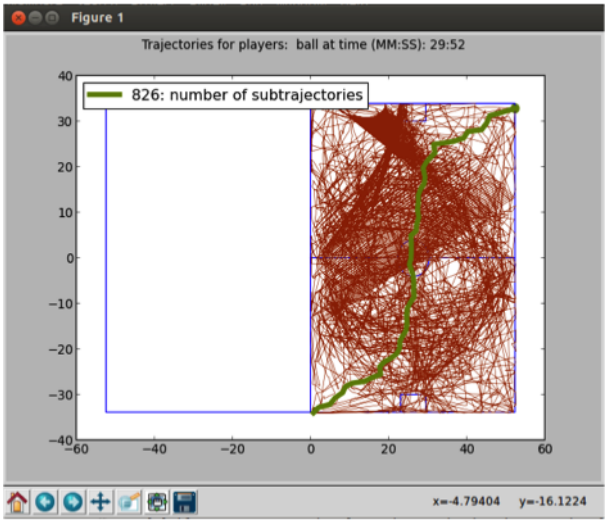
could be based on the analysis of several different matches to understand the nature of the movement.

We also tried to discover the common trajectories of two players. For example we compared players belonging to the same team with same roles (striker) or totally different positions (striker and defender).





What we can see here is that a cluster with a short representative line means that the similar trajectories are very close and the movement happen in almost the same position of the field meaning the similarity is high based on the role and the characteristic we want to study in a player. A long representative line means that the similar trajectory happen in far portions of field. The last experiment we conducted is about the ball movement. From our analysis we can see how we understand, from the clustering, where the play with the ball is mostly developed.



3.5 Speed performance analysis

Having a real time data provided by the sensor with an high frequency one interesting problem we can solve easily is the computing of the speed performance of the player with a very little error. We found interesting know some simple but meaningful statistic as the real-time speed average, the real-time km ran for each player and the real-time performance quantified by expressiveness value as: (stop <= 1km/h , trot <= 11 km/h, low <= 14 km/h , medium <= 17 km/h, high <= 24 km/h, sprint > 24 km/h) but we found also interesting discovery some parameter coming from the comparison between the players.

One statistic value very expressive to understand the difference in terms of running performance is the variance of the speed. Unfortunately, to compute the variance players data is needed and if the number of data recorded becomes huge (we have a recording frequency of 200 HZ) is practically impossible to maintain them in the main memory. A good news is that here we can exploit the streaming algorithm proposed in [3]. In the first section of the paper, is proposed an approach to maintain the variance over a sliding window with a space lower bound of (1/e log N (log N + log R)) bits for(with error at most e) maintaining the sum, where N is the sliding window size and each data value is at most R. (Assuming R = poly(N)).

3.5.1 *The streaming algorithm for the variance*

To compute an estimate of the variance with relative error at most e (with e in the range 0..1). The elements in the data stream are partitioned into buckets by the algorithm. For each bucket Bi , besides maintaining the timestamp ti of the most recent data element in that bucket, the algorithm maintains the following summary information: the number of elements in the bucket (ni), the mean of the elements in the bucket (i), and the variance of the elements in the bucket (Vi). The actual data elements that are assigned to a bucket are not stored. In addition to the buckets maintained by the algorithm, another set of suffix buckets, denoted $B_1 \cdots B_j$, are stored and they represent suffixes of the data stream. Bucket Bi represents all elements in the data stream that arrived after the elements of bucket Bi , that is, Bi = union of Bl with l starting from 1 to i-1 . Except for the bucket Bm , which represents all points arriving after the oldest non-expired bucket, these suffix buckets are not maintained by the algorithm, though their statistics are calculated temporarily during the course of the algorithm.

How to compute the variance? Let $B_1 \cdots B_m$ be the set of histogram buckets at time t. Let B_m be the oldest active bucket. It contains some active elements, including the one with timestamp N, but may also contain expired data. The algorithm maintains statistics corresponding to $B_m\star$, the suffix bucket containing all elements that arrived after bucket B_m. To this end, we use the combination rule for every new data element that arrives. Whenever the oldest bucket gets deleted, we can find the new $B_m\star$ by deleting the contribution of the new oldest active bucket (B_m) from the previous $B_m\star$, using the combination rule. Let $B_m\wr$ refer to the non-expired portion of the bucket B_m , i.e., the set of elements in B_m that are still active. Since we do not know the statistics $B_m\wr$, m and Vm corresponding to bucket $B_m\wr$, we estimate them as follows: nm EST= N + 1 tm; m EST = m ; Vm EST = Vm/2

The statistics for $B_m\wr$ and $B_m\star$ are sufficient to accurately compute the variance at the time instant t. In fact the variance is nothing but the variance corresponding to the bucket B m ,m* obtained by combining $B_m\wr$ and $B_m\star$.

Therefore, by the combination rule, the actual variance ($VAR(t)$) for the current active window is given by:

$$VAR(t) = Vm+Vm*+((nm*nm*)/(nm+nm*))*(mESTm*)^2$$

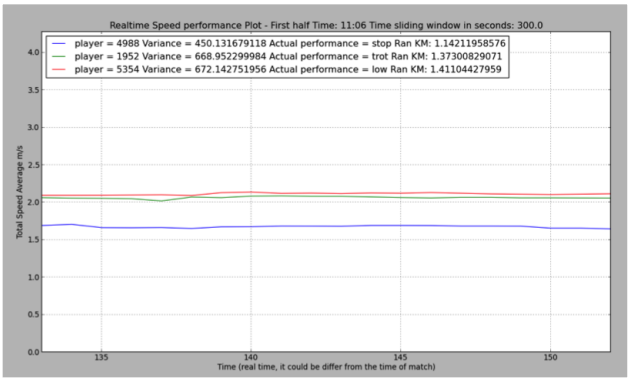


Figure 8: Here we can see the variance of speed in the last 5 minutes (the size of window) respect the total average. Attached to the label of the plot we can see the instantaneous statistic of the performance all of them computed in O(1) thanks to the data come from the sensor.

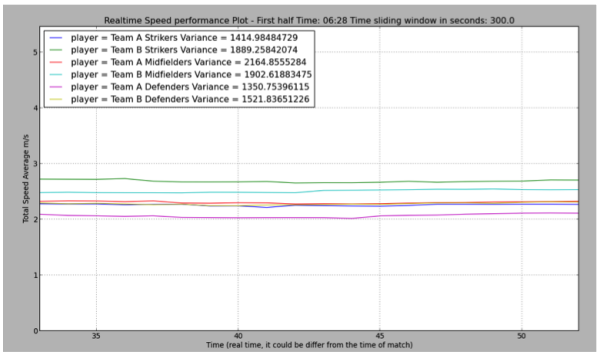


Figure 9: Here instead we can see the aggregate statistic for the player divided by role.

Having the possibility to compute the variance of the speed with a streaming fashion we found interesting to have 2 measure of comparison. The self-comparison for each player, in other word, having at real-time the variance of the player in the last m minutes (the size of the window is decided by the user). The second measure concerning the variance of speed for the average speed of the player divided by the role (only for the movement player: defenders, mid fielders, strikers).

3.5.2 Results of the performance speed analysis

At the end we have the possibility to show in real time such plots. Two examples are reported in this paper.

4. FUTURE WORKS

I/O dependence Our code is high I/O depedent so, in order to speed up the whole operations it would be interesting to use a scalable framework as Map/Reduce and split the operations between many machines.

Incremental trajectories It would be interesting to apply an incremental trajectory clustering [18]

Scalability We could focus a future work comparing the possible different results obtained by applying this framework to different matches.

5. CONCLUSIONS

In this work, we have proposed a new approach to extract hidden informations from a football match dataset yield by an innovative sensor system provided by Fraunhofer. This system in an alternative to video-motion analysis which provides reliable informations and it is a possible new fashon which can be adopted from different clubs and FIFA organization.

The intense competitive schedules of elite soccer clubs require data to be available usually within 24-36 hours postmatch [motion analysis citation] and our system provides all the qualitative data in real-time. Moreover the interesting fact is the scalable nature of our approaches. According to the needs of the football professionals we can analyze other several factor exploiting the techniques coming from the Massive Data Analytics research field with optimal usage of the computing resource (e.g.. the memory usage with the increasing of the data).

When Bill James and the SABR-metricians were making their contributions to baseball, they worked with public data. Now that sport analytics seems to be mainstream, we hope that a new soccer big data era is coming. For this reason we provided a framework of tools which permits the coaches to base their new strategies not only on their intuitions and judgments but also on comprensive (and maybe not human-eye-visible) statistics. Moreover we made it free and public at this URL `http://git.io/ms-8XA`.

6. REFERENCES

[1] B. L. Aleksandar Jankovic and M. Pasic. Analysis of efficient attacks in the 2008 european football championship. *Fizika kultura*, 2, 2009.

[2] D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[3] B. Babcock, M. Datar, R. Motwani, and L. O'Callaghan. Maintaining variance and k-medians over data stream windows. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 234–243. ACM, 2003.

[4] Bizpunk. 2013 mit sloan sports analytics conference, Mar. 2013.

[5] K. Choi and Y. Seo. Automatic initialization for 3d soccer player tracking. *Pattern Recognition Letters*, 32(9):1274 – 1282, 2011.

[6] DEBS2013. The 7th acm international conference on distributed event-based systems.

[7] DEBS2013. The acm debs 2013 grand challenge.

[8] G. dello Sport. Serie a: 1.630 milioni di debiti. club in crisi, i costi non calano. napoli e lazio in attivo, Mar. 2013.

[9] J. Duch, J. S. Waitzman, and L. A. N. Amaral. Quantifying the performance of individual players in a team activity. *PloS one*, 5(6):e10937, 2010.

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

[11] E. W. Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, 21:768–769, 1965.

[12] D. Grabiner. The sabermetric manifesto. *Baseball SABR. Web: http://baseball1. com/bb-data/grabiner/manifesto. html GARGANTA, J. & PINTO, J.. O ensino do Futebol. In A Graça & J. Oliveira (Orgs). O ensino dos jogos desportivos coletivos*, pages 95–135, 1999.

[13] R. Guimera, B. Uzzi, J. Spiro, and L. A. N. Amaral. Team assembly mechanisms determine collaboration network structure and team performance. *Science*, 308(5722):697–702, 2005.

[14] N. James. The role of notational analysis in soccer coaching. *International Journal of Sports Science and Coaching*, 1(2):185–198, 2006.

[15] S. D. Katzenback JR. *The Wisdom of Teams*. Harper Business, 1993.

[16] C. Lago-Peñas, J. Lago-Ballesteros, A. Dellal, and M. Gómez. Game-related statistics that discriminated winning, drawing and losing teams from the spanish soccer league. *Journal of Sports Science and Medicine*, 9(2):288–293, 2010.

[17] J.-G. Lee, J. Han, and K.-Y. Whang. Trajectory clustering: a partition-and-group framework. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pages 593–604. ACM, 2007.

[18] Z. Li, J.-G. Lee, X. Li, and J. Han. Incremental clustering for trajectories. In *Database Systems for Advanced Applications*, pages 32–46. Springer, 2010.

[19] M. Peliš. Sport–kulturní fenomén. *Telesná výchova a šport v kultúre spoločnosti. Fakulta tělesnej výchovy a športu, Univerzita Komenského, Bratislava*, pages 142–148, 2003.

[20] N. Y. Redbulls. Armchair analyst: What we've learned from opta.

[21] W. A. Reilly T. *Science and soccer*. Routledge, 2003.

[22] J. VAN DIERENDONCK. Collaboration: Group theory.

[23] Wikipedia. Progression of british football transfer fee record.

[24] S. Wuchty, B. F. Jones, and B. Uzzi. The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039, 2007.