A Pass Prediction System Derived From The Broadcasting Soccer Video

By

Samriddha Sanyal

Electronics and Communication Sciences Unit
Indian Statistical Institute

Advisor

Prof. Dipti Prasad Mukherjee

Electronics and Communication Sciences Unit
Indian Statistical Institute

A thesis submitted to the Indian Statistical Institute
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computer Science

September 2022

To my boromama late Dwipendranath Talapatra
Ex director, Regional Meteorological Centre, Alipore
and mami Kuhu Talapatra.

Acknowledgements

I am grateful to the institute for the support I have received. I am blessed to have Prof. Dipti Prasad Mukherjee as my advisor. I acknowledge his relentless help and support towards the accomplishment. I appreciate the valuable interactions I have had over the years with all other colleagues at the unit.

I am grateful to my wife Swati for the unwavering love and care throughout my journey. Thank you to my friends: Rakesh, Sandip, Avishek, Aparajita, Sreeya, Koustav da, Suvrojyoti, Suman and others. I can't thank you all enough.

Abstract

In soccer, the most frequent event that occurs is a pass. For a trained eye, there are a myriad of adjectives which could describe the possible pass (e.g., “majestic pass”, “conservative” to “poor-ball”). During the game, defending players constantly try to predict the pass of the attacking player to prevent a goal. So, pass prediction is an important facet to anticipate the game strategy of the participating teams. Related work in this area relies on the information extracted from multiple camera installations in the stadium or data coded by human annotators sitting in the stadium. Aberrating the state-of-the-art, we present a pass prediction system directly derived from the broadcasting video. In order to develop the pass prediction system, three components are needed as follows.

The first component is to track the ball in a given broadcasting video. The problem of tracking ball in a soccer video is challenging because of sudden change in speed and orientation of the soccer ball. Successful tracking in such a scenario depends on the ability of the algorithm to balance prior constraints continuously against the evidence garnered from the sequences of images. In this thesis we propose a particle filter based algorithm that tracks the ball when it changes its direction suddenly or takes high speed. Exact, deterministic tracking algorithms based on discretized functional, suffer from severe limitations in the form of prior constraints. In contrast, our tracking algorithm exploits a probabilistic framework and has shown excellent result even for partial occlusion.

A holy grail for sports analytics is the top-view visualization of the game. The top-view visualization provides the actual between-player distances as opposed to the between-player distances calculated from the side and/or oblique view of a match as shown in the broadcasting video. Therefore the second component of the pass prediction system is the top-view visualization of the match. In this work, we present a factor theory based approach to derive the top-view visualization of the game from the broadcasting sports video. We theoretically prove that the proposed factor theory based approach for top-view visualization is more efficient than the state-of-the-art approach. In addition, as per the proposed approach, we present a model for the top-view visualization by transforming the broadcasting video into a single and static camera visualization. In order to generate the single-camera visualization, the view of the entire ground is needed which is expressed as a solution to a convex optimization function, devised to explore putative matrix completions.

Finally we present a probabilistic framework for pass prediction. The proposed framework predicts pass recipients by integrating two dependent models, designed from the coordinates of the players in abstract top-view visualization. The contribution of the work are generation of the proximity model based on the positions of the opponent team players, and generation of pass region model that is influenced by the concentration of the players of the team who is in possession of the ball. To evaluate the real time efficacy of the proposed pass prediction system, a soccer data set has been introduced. The proposed pass prediction system is compared against recent methods and the ground truth available in the soccer data set. The proposed method outperforms the existing approaches by a noticeable margin.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Analysis of soccer match

Analysis of soccer match from the broadcast video of the game captures complex movement of the players and between-player interactions with respect to various situations that occur in a soccer match. This analysis of soccer match includes analysis of soccer ball possession by a team, analysis of passing of ball from one player to his teammate and above all designing a winning strategy for a team. This thesis intends to address some of these issues related to soccer analytics using computer vision and machine learning based approaches.

The role of a smart visual analytic system for soccer is complimentary in nature. It adds to the expertise of a soccer coach or analyst. There are a number of challenges to automatically derive meaningful observation in a soccer match by a visual analytic system. This is because the movement patterns of the players in a match vary from a smooth regular geometric pattern to a more chaotic form. The direction and rolling speed of a soccer ball vary significantly. The broadcast video often looks at the region of interest in a running soccer match ignoring off-the-ball players who often change the focus of the game. The skill set, talent and the ability of players to influence a game pose varying level of challenges to a soccer visual analytic system.

In the related context, one of the earliest attempts is to integrate soccer related domain knowledge in a soccer match analysis system. This is described next.

1.2 Domain knowledge for soccer analysis

The insights into a match by a soccer analyst is called domain knowledge [Stein et al., 2013]. The domain knowledge integration can enhance the quality of soccer match analysis. Therefore incorporation of domain knowledge into visual analytic system is an immediate choice. To that end, [Lucey et al., 2014] developed a method to estimate the probability of winning in a soccer match with the help of domain knowledge. The key finding of [Lucey et al., 2014] is besides the special events (e.g. free kick, foul), there are other factors like scoring goals, disrupting a pass etc. that influence the probability of winning. [Rathod and Nikam, 2014] proposed a fuzzy inference system to reveal interesting events like goal, pass-break etc. of a match.

[Xu et al., 2008] proposed a model to derive the domain knowledge from video and commentary by employing a conditional random field model. [Perin et al., 2013] introduced a software interface to derive insights by analyzing soccer data. Their interface provides a series of connected visual representations of the whole game. Furthermore, [Perin et al., 2013] argued that a quantitative or statistical analysis is not sufficient to understand the performance of a team. Analysis of collective movement of the players in the ground may be useful to develop a better visual analytic model.

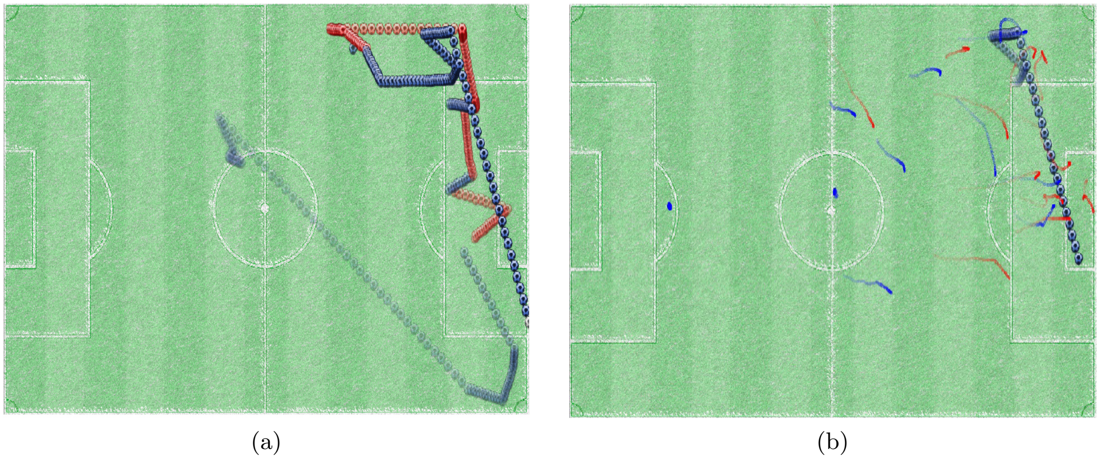


Figure 1.1: A glimpse of the visualization of the game presented by [Stein et al., 2015]. (a) A visualization of the trajectory of the ball. (b) The colors of the ball with trajectory representing the ball-possessing teams (colors red and blue represent two teams).

The term collective movement is used to describe a similar, coordinated and interdependent movement of a group of players. The analysis of collective movement requires the exact recording of each individual movement. The resulting trajectory information often consists of coordinates of each individuals. In soccer, players of a team want to reach a goal collectively. Players need to take decisions and derive strategies in cooperation with their team members as well as in competition with the players of the opposing team. Towards this end [Stein et al., 2019] presented a visual search system to analyse the collective movement in a soccer match. [Reep and Benjamin, 1968] and [Yue et al., 2014] used match statistics such as ball possession, the number of shots on the opposing goal, or the number of won tackles to analyse the collective movement of the participating teams. Identifying interesting game situations of the players from the broadcasting video of a match is also an important component of the visual analytic system.

The interesting game situation is defined as a set of features derived from the match visualization that describes the situation of a match with respect to time. Considering movement features as a function of time, [Stein et al., 2015] presented a visual analytic system for identifying interesting situations in a match dependent on the movement of the players as shown in Fig. 1.1. Their model is inspired by [Von Landesberger et al., 2014] who have used time series to find situations significantly different from the previous movement behavior with respect to time. [Sacha et al., 2014] presented a system for the analysis of position-based soccer data. The system is interactive and depends on the classification of interesting events.

Analysis of player trajectory may be useful to derive a visual analytic model.

The player trajectory is defined as the path approximated from the co-ordinate information of the respective player in the previous frames. In this regard [Sacha et al., 2017] presented a dynamic approach for trajectory visualization of a player by combining the co-ordinate information and clustering techniques. However, the uncertainty arises while the cluster of players is involved in trajectory approximation as the movement of the player cluster follows chaotic movement. [Janetzko and Fabrikant, 2017] proposed visualization methods to depict uncertainty in the trajectory of the cluster representatives of a team. To make visual analytic system more efficient, several match statistics can be incorporated.

The soccer match statistics are derived from the events of the game that describe the quality and performance of the teams. The match statistics consists of ball possession, pass prediction etc. The match statistics with a visual analytic system are available for video games like FIFA 16 [Cotta et al., 2016]. What if the same tools are available during broadcasting real time video of a soccer match? We continue discussion on several open problems in the next section.

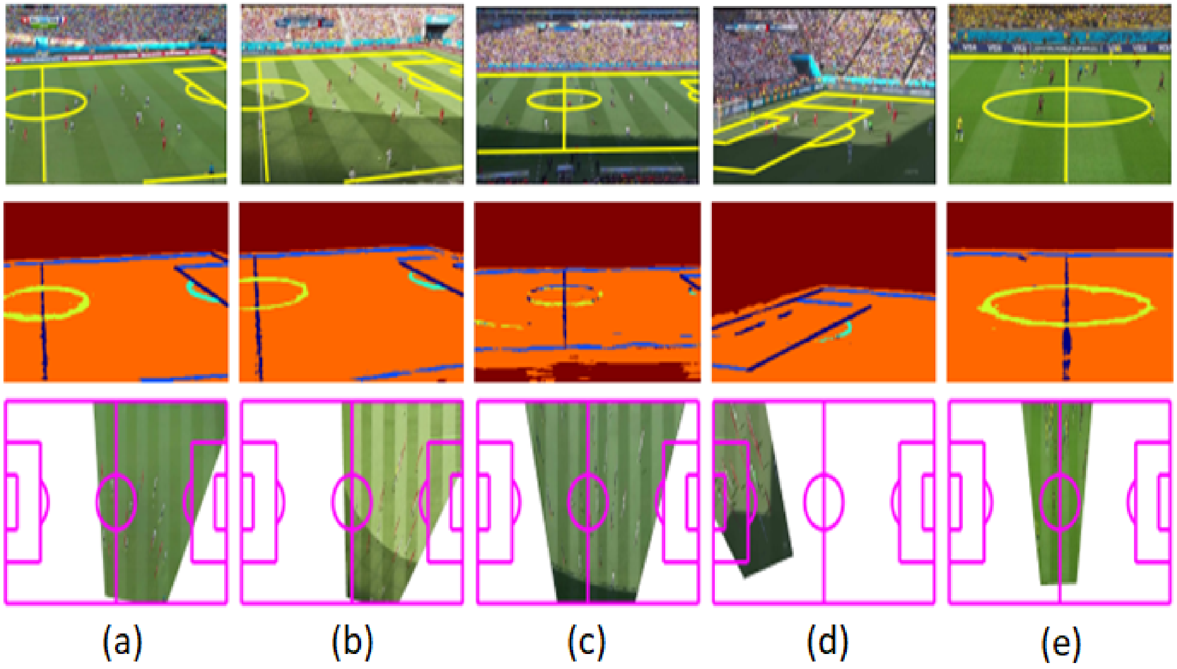


Figure 1.2: (a)-(e): are the examples of the top-view visualization problem proposed by [Homayounfar et al., 2017]. The first row is the given images from a soccer match video. In the second row, lines and center circle are extracted. In the third row, the given images of soccer match are registered inside the top-view model with the help of the extracted lines and circle.

1.3 Open problems in soccer analytics

1.3.1 Top-view visualization

Broadcasting video of a soccer match provides 2D object to image projection of a 3D event, in this case a soccer match. Therefore the correct distance between any two players can not be realized from the broadcasting video. To recover the correct between-player distance a top-view visualization of the match is always helpful. The top-view visualization may be assumed as the display of the game by an imaginary camera installed at the top of the center position of the ground. [Homayounfar et al., 2017] derived the top-view visualization of a set of given images of football match as shown in Fig. 1.2. [Sha et al., 2020] improved the method and derived a better visualization. However, given the multi-camera setting typical to a live telecast, deriving the top-view visualization of a full match from the broadcast video is still a challenging open problem.

1.3.2 Ball-possession statistics

Earlier the ball-possession statistics of a team is defined as the amount of time it controls the ball during a match. The controversy arises with the definition. Is the amount of time really the best basis for this statistic? After all, if a goalkeeper controls the ball for a few seconds, without doing anything specific with it, can it work in any way for the benefit of his team? Therefore, the ball-possession statistic of a team is redefined by the number of valid passes made by the team divided by the total number of valid passes made by both the teams. A network flow analysis is utilized by [Sarkar et al., 2019] to generate ball-possession statistic. The relationship between the match outcome and ball possession in UEFA champions league is studied by [Farias et al., 2020]. However, deriving the game strategy with the help of ball-possession statistics is an open problem.

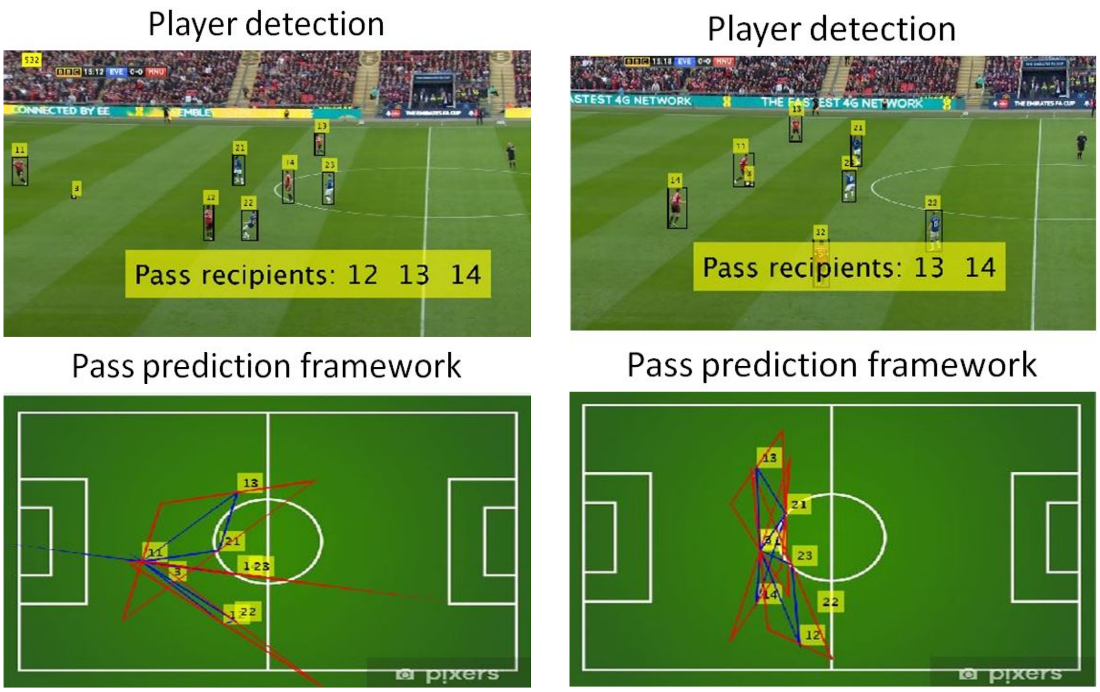
1.3.3 Team formation analysis

In soccer, [Rein and Memmert, 2016] claimed that team tactics are mainly determined by team formations on the pitch. Specifically, team formations can be characterized by the roles and places of players on the ground. Therefore, team formations influence the team performances inevitably in various aspects, such as defining the roles of the players, increasing coverage and determining strike strategies. Towards that end, [Bialkowski et al., 2014a,b] used a minimum entropy data partitioning method to detect formations from positions of individual players and identified different patterns of team behaviors. [Wei et al., 2013] discovered the top offensive and defensive plays applied in a game based on detected team formations. [Wu et al., 2018] proposed a visual analytic system for a comprehensive investigation of team formations. They develop a set of interactive labeling tools to complement the automatic methods. However the proposed model is semi-automatic and depends on human intervention for initial labeling. Deriving a visual analytic system to analyse the team formation is still an open problem.

1.3.4 Pass network

Considering the players as nodes, accomplished passes as edges [Cotta et al., 2011] analysed the pass network among the players of the Spanish team (the world champion in the FIFA

World Cup 2010), with the objective of explaining the results obtained from the behavior at the network level. However, the work does not give any information about the relation between the pass network and player positions. [Gonçalves et al., 2018] explored how passing networks and positioning variables can be linked to the match outcome in youth elite association football. But the work does not utilise the information where the pass is occurring in the ground. [Kawasaki et al., 2019] proposed a method to create a pass network based on the measurement of the pass positions inside the ground. The pass positions were determined from the player positions measured by the automatic tracking system for soccer players. However, the model is interactive and needs human intervention. Designing a fully automated visual analytic system for pass network is still an open problem.


Player detection
Pass recipients: 12 13 14
Pass prediction framework
Player detection
Pass recipients: 13 14
Pass prediction framework

(a)

(b)

Figure 1.3: Example of the pass prediction model from [Sanyal, 2021]. Tracking of the players and the ball in the given video is shown in the first row. Based on the tracking information, the possible pass recipients are predicted as shown in the second row. (a) and (b) are the first and the last frames of a sample soccer video produced from the model proposed in [Sanyal, 2021].

1.3.5 Pass prediction

The game of soccer involves an act of one team trying to score a goal against the other. During the game, defending players constantly try to predict the pass of the attacking player to prevent a goal. Therefore, pass prediction is an important facet to reveal the game strategy of participating teams. Towards this end, [Spearman et al., 2018] used concepts of interception and control time for pass prediction in soccer. [Felsen et al., 2017] proposed a framework based on the top-view visualization and video frame information for forecasting future events in team sports videos. [Sanyal, 2021] proposed a pass prediction framework by integrating two models dependent on each other. The output of the system proposed by [Sanyal, 2021] is shown in Fig. 1.3. However, image information is not used to predict the possible pass recipients by [Sanyal, 2021]. Therefore integrating an image-based prediction model may give a better visual analytic model for pass prediction. Hence incorporating the image based prediction model is an open problem.

The pass prediction problem plays the key role in problems like pass-networking and calculation of ball-possession statistics. Consequently, the strategies for deriving a visual analytic system for pass prediction becomes important. Next we discuss the existing literature, that focus on the visual analytic system for pass prediction.

1.4 Related works

The game strategy of soccer depends on one team trying to intercept the passes initiated by the other team. Over the last decade, researchers have given significant effort to address the pass-prediction problems. In this thesis we focus on deriving a pass-prediction system. There are three important steps to derive a pass-prediction system.

- A model to track the ball during the soccer match
- A top-view visualization model to derive the correct positions of the players and the ball given given in a broadcast video
- A pass-prediction model to predict the possible pass recipients

In the next few paragraphs, we discuss the major literature related to these three challenges.

[Huang et al., 2008] tracked the soccer ball in a video by integrating segmentation based detection and particle filter based tracking. However, detection sometimes failed and then false positives were tracked in successive video frames because the image of the ball cannot be differentiated from other objects based only on the appearance because of the size and the poor features. [Yu et al., 2004] tried to do ball detection. They tracked the ball based on the position of the players and the hitting sound. This method, however, is not always applicable due to the difficulty faced to extract the sound generated from the targeted object. [Xing et al., 2011] tracked multiple players in a sports video by dual-mode two-way Bayesian inference approach. But the method proposed by [Xing et al., 2011] does not work for small object like soccer ball. Next we discuss different visualization approaches.

The broadcasting video provides us a 2D visualization of a 3D event. Therefore the correct distance between any two players can not be realized from the broadcasting video. The top-view visualization that assumes a camera at the top of the soccer field, provides us the correct between-player distances. The top-view player coordinates are widely used in sports analytic [Sha et al., 2018], [Zhan et al., 2018], [Hobbs et al., 2018]. In field-registration method, the task is to find the homography that can map the 2D field from the video frame to a top-view model. In order to handle top-view visualization problem, several methods proposed by [Kim and Hong, 2000], [Farin et al., 2003], [Wang et al., 2006], [Watanabe et al., 2004] used geometric features of the ground such as lines and/or circles or arc.

There are field-registration approaches based on Hough transform [Yamada et al., 2002] and RANSAC [Carr et al., 2012]. [Carr et al., 2012], [Ghanem et al., 2012], [Gupta et al., 2011] developed models for field registration find correspondences between points and line segment of the given image and top-view model. [Okuma et al., 2004] derived the top-view visualization using the initial homography and frame to frame geometric transformation. [Homayounfar et al., 2017] extended the works of [Hayet et al., 2004] and [Hayet and Piater, 2007] with CNN-based semantic detection to accurately address the top-view visualization problem. However, deriving the top-view visualization of a full match from the broadcasting video is an open problem.

Recently, research related to ball-passing becomes popular to analyze football match strategy. The attributes of a pass has been used to understand how the team performance may be optimized [Duch et al., 2010], [Lusher et al., 2010], [Grund, 2012] and [Fewell et al., 2012]. [Cintia et al., 2015] showed how the player-tracking information play an important role to accurately predict a successful pass. Indeed, the authors demonstrated that a view on football passing data has the potential of revealing hidden behaviours and patterns of the players. [López Peña and Touchette, 2012] demonstrated that winning teams have complex ball-passing dynamics between teammates. The insight helps to identify the key players involved during the offense and to describe the relationship within team positioning strategy at different levels of analysis [Gama et al., 2014], [Clemente et al., 2014]. [Link and Hoernig, 2017] proposed models for estimating ball possession of individuals and a team based on position data of ball and players. [Chawla et al., 2017b] proposed a model to evaluate quality of a pass in a soccer match.

[Gonçalves et al., 2018] explored how passing networks and positioning variables can be linked to the match outcome in a soccer match. [Goes et al., 2019] proposed a data driven model to quantify pass effectiveness by means of tracking data. The idea is to investigate how a pass disrupts the opposing defense. [Power et al., 2017] used a model based on an objective quantification of the risk and reward of a pass to assess the pass quality. [Rein et al., 2017] quantified pass effectiveness using Voronoi diagrams [Taki et al., 1996]. [Chawla et al., 2017a] proposed a system to automatically classify the pass on the field without linking it to goal scoring.

[Spearman et al., 2018] used physical concepts of interception and control time for pass prediction in soccer. However, their model is based on control time and velocity of players. Naturally, the model fails due to abrupt velocity change of players. [Felsen et al., 2017] proposed a framework for forecasting future events in team sports videos directly from visual inputs. They predict pass from the visualization of top-view of the game for water polo and basketball dataset. However, their framework does not incorporate contextual knowledge of the game. Incorporating the contextual knowledge in pass-prediction is an important open problem. Next we discuss the aim and contributions of this thesis.

1.5 Contributions of the thesis

In this thesis, we aim to address the pass prediction problem. We develop a visual analytic model that predicts the possible pass recipient directly from the broadcasting football video. The contributions of the thesis are as follows:

- A particle filter based algorithm is developed for tracking small object like soccer ball in a broadcasting video. The proposed algorithm addresses the issues of the state-of-the-art tracking algorithm while tracking smaller object like the ball.

- Inspired by factor theory, a top-view visualization approach is proposed to derive the top-view of a soccer match directly from the broadcasting video. We theoretically prove that the proposed approach for top-view visualization is better than the state-of-the-art approach.
- To implement the top-view visualization model the global view of the ground is necessary. To obtain the global view, a matrix completion paradigm has been introduced.
- A pass prediction model is introduced that predicts possible pass recipients by exploiting the domain knowledge of the game.
- To extract the domain knowledge of the game, prior-free and interdependent models are introduced that exploits the arrangement of the players.
- A theoretical framework to combine inter dependent models is introduced to infer pass-prediction.

1.6 Organisation of the thesis

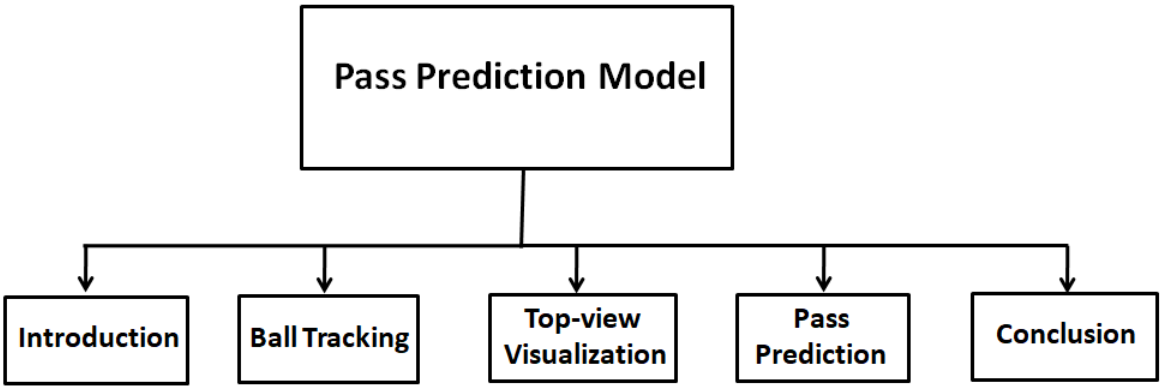


Figure 1.4: Layout of the thesis

The proposed pass prediction system comprises of three modules namely tracking model, top-view visualization model and pass prediction model as shown in Fig. 1.4.

The chapters 2, 3 and 4 of the thesis are dedicated to describe the components of the pass prediction model. In chapter 2, a ball tracking algorithm is developed. The algorithm is based on particle filter and tracks the ball even when it changes its direction suddenly or takes high speed.

In chapter 3, we propose a factor-theory inspired top-view visualization approach with theoretically proven efficacy against the state-of-the-art approach. Furthermore as per the proposed approach, a top-view visualization model has been implemented. In order to design the top-view model, the global view of the ground is necessary. In order to derive the global view of the ground from bits and pieces of ground information available in the video frames, a matrix completion paradigm has been introduced. Challenges originated during the derivation of the top-view model are presented at the end of the chapter.

Finally, empowered with the tracking information from the broadcasting video and the top-view player coordinates, a pass-prediction model is proposed in chapter 4 of the thesis.

We conclude the thesis in chapter 5, where we summarize our contributions and point to the future directions of research.

Chapter 2

Tracking the ball

There are limited applications of computer vision techniques for analysis of sports video [Moeslund et al., 2015]. As a part of our initiative to derive an analytic system from broadcasting video, in this chapter we present a technique to track soccer ball . One major challenge for tracking soccer ball is sudden occlusion by players and markings of the field. The other major issue is when the ball is being kicked or passed by a player [Yu et al., 2004]. The ball can suddenly change its direction of motion. The between-frame motion of the ball can be significant. Most of the conventional trackers fail in such situations. The high speed of ball movement often blurs the image of the ball causing failure of the tracking proposal.

[Pallavi et al., 2008a] proposed a ball tracking method using static and dynamic features. [Huang et al., 2008] computed ball position in a frame by integrating segmentation based detection and particle filter based tracking. However, detection sometimes fails and then false positive is tracked in successive frames because imaged ball cannot be differentiated from other objects based only on the appearance because of size and the poor features. [Yu et al., 2004] tried to detect ball based on the position of the players and the ball hitting sound. [Xing et al., 2011] tracked multiple players in a sports video by Bayesian inference approach. [Pallavi et al., 2008b] proposed a graph based multiple-player detection and tracking approach.

Given the history of soccer ball positions in a video, proposals for possible location of the soccer ball in the next frame may be generated. The proposals are made based on a concept similar to particle filters. We then propose a strategy to select the winner among the different proposals. The proposal is based on the features collected from the ball, namely, the color, edge gradient and shape measure. An iterative scheme weighs each proposal based on the likelihood of the proposal to be a ball in the next frame. This is similar to edge tracking framework proposed in [Pérez et al., 2001].

Our soccer ball tracking algorithm is comprised of the following four steps: (a) Prediction of a region called the *measurement space* around the ball that captures the possible future locations where the ball can be. (b) Weight assignment to points on the measurement space and (c) re-sampling of the winner points from the measurement space based on their weights, (d) prediction of center and radius of the best fitted circle considering the winner points. The predicted best fitted circle represents the ball in the $(i + 1)$ th frame. The block diagram of the entire approach is shown in Fig. 2.1. Finally we spot the ball in the next frame based on the resampled points. The prediction step defines a dynamic model that generates a set of points on the image. These points may lie on the edge of the ball in the frame where we are searching. Next we present the prediction step.

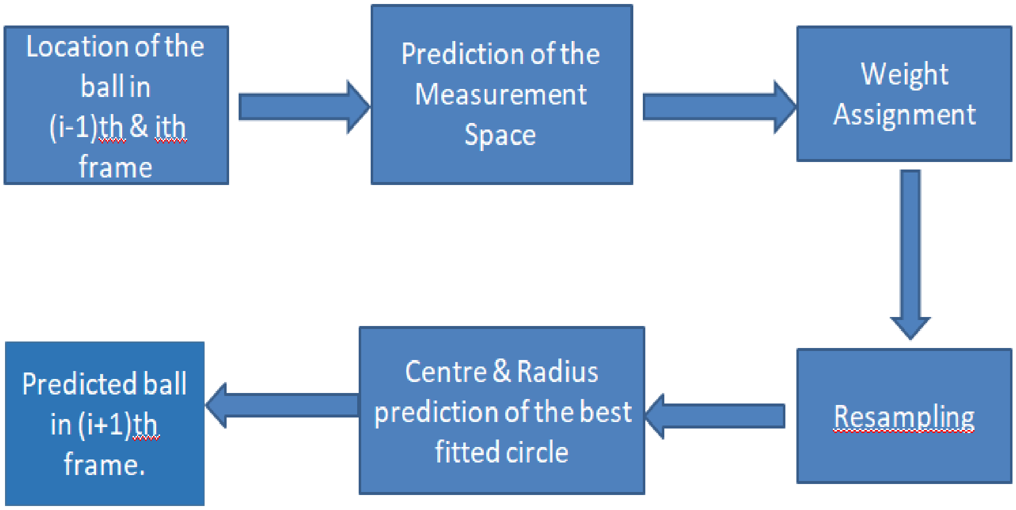

Location of the ball in (i-1)th & ith frame
Prediction of the Measurement Space
Weight Assignment
Resampling
Centre & Radius prediction of the best fitted circle
Predicted ball in (i+1)th frame.

Figure 2.1: Block diagram of the proposed soccer ball tracking algorithm. Based on the position information of the ball given in $(i-1)$th and ith frames respectively, the algorithm predicts the position of the ball in $(i+1)$th frame.

2.1 Prediction

The ball tracking problem is posed as a bi-level pixel tracking problem. At the first level, given previous frames, specifically frame $(i-1)$ and i, the target is to generate a set of shapes mimicking the shape of a ball at frame $(i+1)$. At the second level, the winner shape among the predicted shapes at frame $(i+1)$ is identified. The properties of image pixels at frame $(i+1)$ representing the potential ball shape should ideally support the winner shape at frame $(i+1)$. Based on the above ideas, we propose a probabilistic framework to track the ball.

Assume the ball and the center of the ball are X_i and O_i in frame i respectively. We initialize two points as the center of the ball and another one is on the circumference of the ball. The radius of the ball r is the distance between two initialization points. The circular shape X_i of each predicted ball position may be specified using a set of k number of discrete points $(r\cos\theta, r\sin\theta)$ on the edge of the ball. For k discrete points, k number of equally spaced θ values are discretized between $[0, 2\pi]$ on the perimeter of the ball.

The entire scheme involving video frames $(i-1)$ to $(i+1)$ is shown in Fig. 2.2. The absolute distance travelled by the ball between O_{i-1} and O_i in frames $(i-1)$ and i is d. Let us assume that the ball is visible in the $(i+1)$th frame such that given video frame dimension $(row \times col)$, $d \leq \min(\frac{row}{2}, \frac{col}{2})$. Using the motion of the ball, we construct a space called measurement space that will contain all the possible locations of the ball at the $(i+1)$th frame. In order to construct the measurement space, a two-step transformation T is applied on the X_i. First, X_i is translated along the x axis by $(d, 0)$ as shown in Fig. 2.2(c). Second the translated X_i is rotated around O_i further with radius d as shown in Fig. 2.2(d). Thus for any $v \in X_i$, the T can be written as:

$$T(v) = v', \tag{2.1}$$

resulting points v' in measurement space.

In order to characterize the possible ball locations of the measurement space, we consider the rotation angle θ_b of the X_i involved in the transformation T. For each such θ_b, a configuration matrix M^{θ_b} of dimension $(k \times 2)$ is constructed that contains all the k representative points of the rotated circle X_i described in Fig. 2.2(d). Next we propose a tracking framework to predict the winner shape X_{i+1} at frame $(i+1)$.

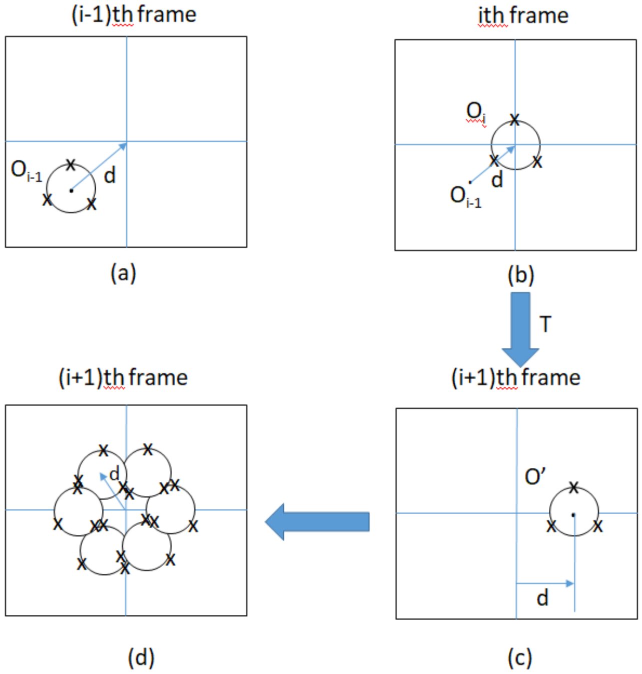


Figure 2.2: (a) and (b): The center of the ball O_{i-1} and O_i in $(i-1)$ and ith frame respectively. (c) The position of the ball in frame $(i+1)$ after translation by $(d, 0)$. The translated circle is at center O'. (d) The *measurement space* constructed by rotating the translated circle with center at $O^{'}$ with radius d.

2.2 Tracking framework

The tracking scheme predicts X_{i+1} based on two factors. First, the dynamics in the expected position of the ball (which is represented using k discrete points on the edge of the ball) is modeled using a kernel density function. The density function is a second order Markov model defined using q as follows.

$$p(X_{i+1}|X_{0:i}) = q(X_{i+1}; X_{i-1:i}), \forall i \geq 2. \tag{2.2}$$

The probability of X_{i+1} given all the previous frames starting the 0th frame, $p(X_{i+1}|X_{0:i})$, is assumed to be dependent on previous two frames i and $(i-1)$ for the second order model. The second factor is the position of the ball recovered using a resampling method from a set of points likely to lie on the ball. In the subsequent section, the exact form of q is defined.

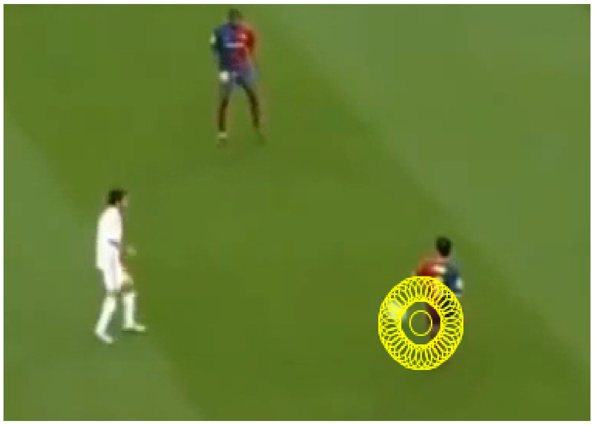


Figure 2.3: The collection of yellow circles shown in the frame is an example of measurement space.

2.2.1 Second order dynamics

The purpose of q is to predict points X_{i+1} from a distribution. The k points in X_{i+1} may be represented as a collection of point x_j, $j = 1, 2, \ldots, k$, $X_{i+1} = \{x_1, x_2, \ldots, x_k\}$. The centre of the ball in frame $(i+1)$ is O_{i+1}. The dynamics in predicting the point x_j can be modeled following Eq. 2.1 when the rotation angle θ_b can be sampled from a zero-mean Gaussian distribution. The angular variance for the zero-mean Gaussian distribution can be measured as variance of the radius d in the previous video frame i. At the same time, the position of the ball may vary abruptly due to sudden but significant between-frame change in the velocity of the ball. Therefore, following [Pérez et al., 2001], we propose a weighted convex combination of uniform distribution and normal distribution as mentioned earlier.

$$q(\theta_b) = \frac{\nu}{\pi} + (1-\nu)N(0, \theta_b), \tag{2.3}$$

where ν is an user given parameter between $[0, 1]$. Essentially q captures the sudden direction change of the ball. We assume the given information (centers and radii of all circles those are likely to contain the ball) up to n frames are mutually independent and follow second order Markov model. Then $p(X_{0:n})$ satisfies the assumption $p(X_{0:n}) = p(X_{0:1}) \prod_{i=2}^{n} q(X_i; X_{i-2:i-1})$ that makes the prior useful to compute the probability of tracking the ball given the prior information. Next we utilize the image based data model to compute the probability of getting the ball given prior information.

2.2.2 On/off the ball

The probability that a predicted contour is in the vicinity of the true contour containing the ball given all the prior information is defined as the data model. The data model is represented as $p(Y(u)|X_{0:n})$, where u is a discritized point representation in the image space from universal configuration matrix Ω. Y represents collection of features derived from pixel at image location u. Consider the universal configuration matrix set as $\Omega = \{u : u \in \cup_{\theta_b} M_{(i+1)}^{\theta_b}\}$ as a collection of all the configuration matrix $M_{(i+1)}^{\theta_b}$ generated at the given frame $(i+1)$. We define $Y : \Omega \to \mathbb{R}^3$ as $Y(u) = (Y_1(u), Y_2(u), S(u))$, where $Y_1(u) = I(u)$ represents the gray scale value at the location $u \in \Omega$. The representative video frame is defined as a two dimensional image I. $Y_2(u) = |\nabla I(u)|$ is the norm of image gradient at the location u. $S(u)$ is a shape measure proposed in [Dryden and Mardia, 1998]. To define the shape measure, let us consider $M^{\theta_b}_{(i+1)}$ be the configuration matrix that contributes the point u in Ω. Therefore, S is defined as:

$$S(u) = \left\| (u - M^{\bar{\theta_b}}_{(i+1)}) \right\|, \tag{2.4}$$

where $M^{\bar{\theta_b}}_{(i+1)}$ is the centroid of all the row vectors of $M^{\theta_b}_{(i+1)}$. The shape measure S evaluates how likely a point lies on the edge of the ball satisfying the shape of the ball given as prior. We prove that S defined in Eq. 2.4 is a shape measure in Lemma 2.1 as defined by [Dryden and Mardia, 1998].

Lemma 2.1. *S in Eq. (2.4) is a shape measure.*

The proof of the Lemma 2.1 is in appendix A.

Our objective is to compute the probability of a point lying on the edge of the ball given the detailed information of the ball in the previous frames. To compute $p(Y(u)|X_{0:n})$, we assume that:

$$\begin{aligned} p(Y(u)) \quad &\propto [p(Y_1(u)) + p(Y_2(u)) + p(S(u))] \\ &= C[p(Y_1(u)) + p(Y_2(u)) + p(S(u))], \end{aligned}$$

where C is a normalization constant. Following the assumption on $p(Y(u))$, we can write $p(Y(u)|X_{0:n}) = C[p(Y_1(u)|X_{0:n}) + p(Y_2(u)|X_{0:n}) + p(S(u)|X_{0:n})]$. Assuming all the points of Ω are independent, we can write:

$$p(Y(u)|X_{0:n}) = \prod_{u \in \Omega} p(Y(u)|X_{n-1:n}). \tag{2.5}$$

We refer $p(Y(u)|X_{n-1:n})$ in Eq. (2.5) as p_{on}. That is, p_{on} denotes the probability of getting a point on the edge of the ball given the ball locations of the previous two frames. $p(Y(u))$ is denoted as p_{off} and refers to the probability of getting a point on the ball without any prior information. So, we can write Eq. (2.5) as:

$$\begin{aligned} p(Y(u) \mid X_{0:n}) &= p(Y(u)|X_{n-1:n}) \\ &= \prod_{u \in \Omega - X_n} p_{\text{off}}(Y(u)) \prod_{u \in (\Omega \bigcap X_n)} p_{\text{on}}(Y(u)|X_{n-1:n}) \\ &= \prod_{u \in \Omega} p_{\text{off}}(Y(u)) \prod_{u \in (\Omega \bigcap X_n)} \frac{p_{\text{on}}(Y(u)|X_{n-1:n})}{p_{\text{off}}(Y(u))}. \end{aligned} \tag{2.6}$$

Next we discuss the choice of p_{off}.

2.2.3 p_{off} calculation

The p_{off} can be empirically computed using the distribution of the norm of the gradient, pixel value and shape measure of the Ω. In our experiments, these empirical distributions are computed using an exponential distribution. The advantage of exponential distribution is that it reflects the deviation from the true value more clearly than a non-exponential one. For $u \in \Omega$, define $Y(u) = (Y_1(u), Y_2(u), S(u))$ where $Y_1(u) = I(u)$, the pixel value at the location u, $Y_2(u) = \nabla I(u)$, the gradient and $S(u)$ is the shape measure. Therefore the p_{off} is defined as:

$$p_{\text{off}}(u) \propto e^{-\|\nabla I(u)\|} + e^{-I(u)} + e^{-d_1}, \tag{2.7}$$

where d_1 is the distance between u and center of the ball at ith frame. Next we discuss p_{on}.

2.2.4 p_{on} calculation

In order to learn the priori, p_{on} computation is needed. To compute the p_{on}, the probability distribution for the features color, gradient and shape should be assigned. We first discuss the probability distribution assigned for color then shape. We take $\frac{e^{-(d_2^2)/2}}{\sqrt{2\pi}}$ as probability density function for color where d_2 is the Euclidean distance between the gray scale value present in u and the center of the ball. For shape measure we take $e^{-\frac{d_1}{S(u)}}$ where d_1 is the distance between u and center of ball at ith frame. We take exponential distribution for color and shape so that if any point which is far from the ball's original value will get less weight for color and shape. Next, we discuss the gradient on the edge of the ball.

The empirical distribution of the gradient over the contour of interest assumes a wide range of values. This is because the features of the ball are highly variable due to abrupt and high-speed movement. To capture highly variable behavior of the features in an adaptive way, it may be better to keep the data likelihood p_{on} as less informative as possible. To keep the p_{on} less informative, we will use an uniform distribution. Now, our aim is to give two different distribution functions for those points which lie on the edge of the soccer ball and off the ball respectively.

To assign two different distributions, first we need a function $\rho : \Omega \to \{0, 1\}$ that determines whether a point lies on/off the edge of the ball. To define ρ, next we construct an automatically calculated threshold that decides whether a point is on/off the ball.

Suppose we have k number of representative points $\{(x_{i-1}^m, y_{i-1)}^m\}_{m=1}^k$ and $\{(x_i^m, y_i^m)\}_{m=1}^k$ from frames $(i-1)$ and i respectively. We denote gradient value of the corresponding coordinates $\|\nabla I(x, y)\|$. Considering all the given representative points, we define the threshold TH_i for frame i as:

$$\text{TH}_i = 1/k \sum_{m=1}^{k} \big\| \|\nabla I(x_{i-1}^m, y_{i-1}^m)\| - \|\nabla I(x_i^m, y_i^m)\| \big\|. \tag{2.8}$$

Now we discuss how TH_i characterizes the ρ. Suppose (x_i^m, y_i^m) is on the edge of the ball in the ith frame. After applying T on (x_i^m, y_i^m) we get (a, b) in Ω. i.e. $T(x_i^m, y_i^m) = (a, b)$. We define (a, b) is on the edge of the ball or $\rho(a, b) = 1$ if $\|\|\nabla I(a, b)\| - \|\nabla I(x_i^m, y_i^m)\|\| \leq \text{TH}_i$ and 0 otherwise. Next we assign two different distribution functions to all the $u \in \Omega$ lying on or off the ball respectively with the help of ρ.

To assign appropriate distribution to $u \in \Omega$, we consider $\frac{\rho(u)}{\pi} + (1 - \rho(u))N(\psi(u); 0, \frac{\sigma_\psi^2}{|\nabla I(u)|})$, where $\psi(u)$ is the angle variation between $\nabla I(u)^\perp$ and $I(O_i)$. Depending on the value of ρ as 0 or 1 the term $\frac{\rho(u)}{\pi} + (1 - \rho(u))N(\psi(u); 0, \frac{\sigma_\psi^2}{|\nabla I(u)|})$ assigns distribution to the u. Now we compute the probability of getting a point lying on the edge of the ball given prior information i.e. p_{on}. The p_{on} is defined as:

$$p_{\text{on}}((Y_1(u), Y_2(u), S(u))|X_{i-1:i}) \propto \frac{\rho(u)}{\pi} + (1-\rho(u))N(\psi(u); 0, \frac{\sigma_\psi^2}{|\nabla I(u)|}) + \frac{e^{-(d_2^2)/2}}{\sqrt{2\pi}} + e^{-\frac{d_1}{S(u)}}. \tag{2.9}$$

Upto now we have both p_{on} and p_{off}. Next we calculate the likelihood ratio l which is the ratio of p_{on} and p_{off}.

2.2.5 Likelihood ratio

The likelihood ratio l implies how likely a point is on the edge or part of the ball shape. In that case, l is directly proportional to p_{on} and inversely proportional to p_{off}. We consider the likelihood ratio $l = \frac{p_{\text{on}}}{p_{\text{off}}}$ which can be deduced from Eq. Eq. 2.7 and 2.9 upto a multiplicative constant.

Now the question is how likely the predicted circle can circumscribe the real ball based on $Y(u)$. We need to derive $p(X_{0:n}|Y(u))$ because there is no closed form expression of $p(X_{0:n}|Y(u))$ available from the given information. However, $p(X_{0:n}|Y(u))$ satisfies the following lemma 2.2 involving the term l and q.

Lemma 2.2.

$$p(X_{0:n}|Y(u)) \propto p(X_{0:1}) \prod_{i=2}^{n} q(X_i; X_{i-2:i-1}) \prod_{u\epsilon(\Omega \bigcap X_n)} l(Y(u)). \tag{2.10}$$

For proof of the above statement we refer **Appendix** A. Next we compute $p(X_{0:n}|Y(u))$ in an iterative way.

2.2.6 $p(X_{0:i}|Y(u))$ computation

Our ball tracking algorithm is based on recursive computation of posterior densities of interest. Let us consider the posterior densities of interest at the ith frame is $p_i(X_{0:i} \mid Y)$. From lemma 2.2 we construct the following recursive relation:

$$p_i(X_{0:i} \mid Y) \propto p_{i-1}(X_{0:(i-1)}|Y) q(X_i; X_{(i-2):(i-1)}) \prod_{u\in(\Omega \bigcap X_n)} l(Y(u)). \tag{2.11}$$

Though we have analytical expression for l and q but the recursion of Eq. 2.11 cannot be computed analytically. Therefore the idea is to approximate the p_i by a finite number of points. Towards this end, we generate points from the distribution p_{i-1}. The generation of samples from p_{i-1} is then obtained by the following process.

Let us consider each point of Ω is chosen by sampling from the proposal density function f over Ω. Since we are seeking points from the unknown p_i, so we will go for *importance sampling* principle [Arulampalam et al., 2002]. Let $\hat{x}^m$, $m = 1, 2, \ldots, N_s$ be the sample that are generated from a proposal $q(.)$ called importance density. Suppose $p_i(\hat{x})$ is a probability density from which it is difficult to draw samples but $p_i(\hat{x}) \propto \pi(\hat{x})$ and $\pi(\hat{x})$ can be computed.

Then, a weighted approximation to the density p_i is given by:

$$p_i(\hat{x}) \approx \omega_m \sum_{m=1}^{N_s} \delta(\hat{x} - \hat{x}^m), \tag{2.12}$$

where $\sum_{m=1}^{N_s} \omega_m = 1$ and $\omega_m \propto \frac{\pi(\hat{x}^m)}{q(\hat{x}^m)}$. δ is Delta-Dirac measure. Now the question is how to compute the $\pi(\hat{x}^m)$. With the help of Eq. 2.11 of p_i, the ratio $\frac{p_i}{fp_{i-1}}$ can be written as $\frac{ql}{f}$. Now $\pi(\hat{x}^m)$ reads:

$$\pi(\hat{x}^m) \propto \frac{q(\hat{x}^m)l(\hat{x}^m)}{f(\hat{x}^m)}. \tag{2.13}$$

It can be shown that the optimal proposal pdf $f = ql/\int_{X_{n-1}} ql$ [Doucet et al., 2000], whose denominator cannot be computed analytically in our case. The chosen proposal pdf must then be sufficiently close to the optimal one such that the weights do not degenerate (i.e., become extremely small) in the re-weighting process. Next subsection we discuss our resampling strategy to select winning points from measurement space.

2.2.7 Resampling

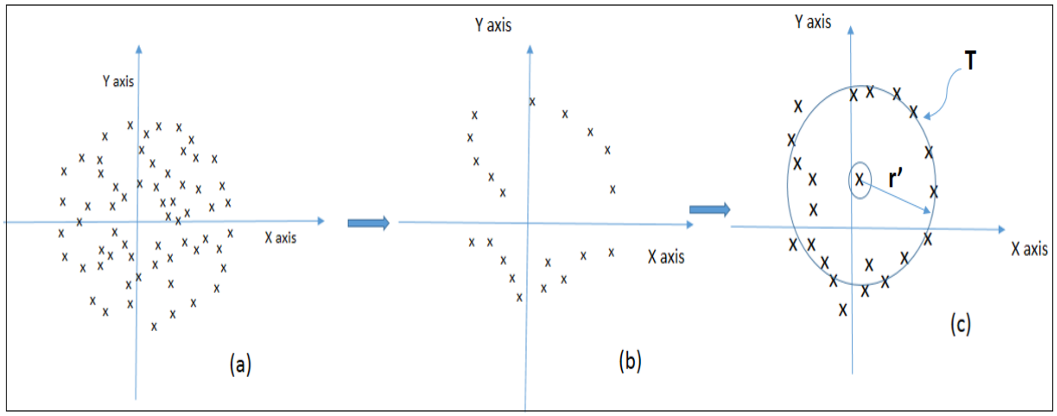


Figure 2.4: Resampling strategy. (a) Measurement space as a collection of points. (b) We resample points from each quadrant based on their likelihood ratio. (c) We calculate weighted mean of all resampled points. The weighted mean is the likely center of the ball.

Let us consider that the universal configuration matrix set (Ω) has total of $N = (k \times \frac{2\pi}{\theta_b})$ points. So, there are total N number of points from which resampling should be done. In Fig. 2.4(a) from each quadrant we resample k points based on the likelihood ratio l as shown in Fig. 2.4(b). In case some quadrant has less than k points then we sample all the points. Note that the total number of resampled points $4k$ is smaller than N. Then we find weighted mean of all $4k$ resampled points, where weights are calculated from Eq. 2.13. To give impact of weight to the center position, we compute weighted centroid $(\bar{x}, \bar{y})$ of the resampled points. From the computed center $(\bar{x}, \bar{y})$ we calculate weighted mean distance $(\bar{r})$ of the resampled points. We take $(\bar{x}, \bar{y})$ as center and $\bar{r}$ as radius then draw the circle as shown Fig. 2.4(c). This is the predicted circle containing the ball of frame $(i + 1)$ based on frame i and $(i - 1)$.

Fig. 2.5 shows an example of resampled points in a soccer video frame.

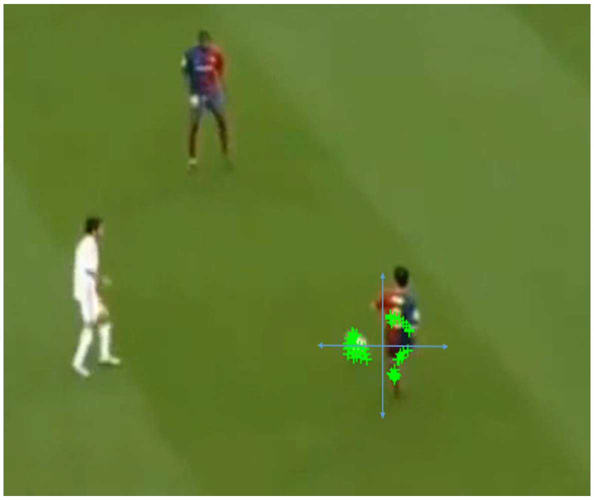


Figure 2.5: An example of resampled points in a video frame.

2.3 Experiment and result

The implementation details of the system and experimental results are presented. The implementation of ball tracking algorithm is given in algorithm 2.1. We first give a short description of the sports video data used in the experiments. Then, we describe parameters used for our football tracking algorithm. Thereafter we evaluate our algorithm with ground truth and compare with color based mean shift non-rigid object tracking algorithm.

2.3.1 Sports video data

One of the motivations of this work is to track football in a soccer match. The video data used for our experiment is directly collected from [Dat, 2022a]. In this dataset one team jersey is white. So, it is difficult for any algorithm to track small white ball based only on color information. Camera movement also creates challenge to track football. It makes ball tracking critical by background subtraction method. We also collect another video dataset [Dat, 2022b]. Here no team has white jersey. We manually label the ground truth of position of the ball. We compare our algorithm with the representative work in [Comaniciu et al., 2000] which deals with the problem of non-rigid object tracking based on mean shift and color histogram.

2.3.2 Parameters

We refer Fig. 2.2(d). We create 36 new circles (we have taken $\theta_b = 10°$) to construct measurement space in each frame. Considering the set of points representing the ball X_i in $(i+1)$th frame, the total number of circles are $36 + 1 = 37$ in $(i+1)$th frame. We take 15 representative points each on X_{i-1} and X_i. Therefore, considering the soccer ball location X_i in the $(i+1)$th frame, the total number of points on all 37 circles potentially representing the ball

Algorithm 2.1 Football tracking algorithm

```
Input:
1. $X_{n-1}$ and $X_n$ represent the soccer ball in $(n-1)$th and $n$th frames respectively.
2. $X_{n-1}$ and $X_n$, each contains $k$ points $(r\cos\theta_j, r\sin\theta_j)$, $j = 1, ..., k$, on the circle perimeter

Step 1 Define measurement space as in section 2.1.
Step 2 Weight particle set $N_s$. Compute $\pi(\hat{x}^m) = K\frac{q(\hat{x}^m)l(\hat{x}^m)}{f(\hat{x}^m)}$ with the normalization constant
       $K$ such that $\sum_{m=1}^{N}\pi^m = 1$.
Step 3 Resample $4k$ particles as defined in section 2.2.7.
Step 4 Draw axis taking center of $X_n$ as origin. At each quadrant, draw top $k$ sample points
       having maximum discrete probability $\{\pi^m\}$ over $\{1, 2, .., N_s\}$ as described in section
       2.2.7.
Step 5 Predict circle representing soccer ball location in $(i+1)$th frame with center $(\bar{x}, \bar{y})$ and
       radius $\bar{r}$ as described in section 2.2.7.
```

in $(i+1)$th frame are $15 \times 37 = 555$. The parameter ν of Eq. 2.3 is 0.05 and $N_s = 570$. The total number of resampled points $4k = 60$.

2.3.3 Quantitative Evaluation

Our main goal is to assess the effectiveness of our proposed algorithm over the state-of-the art algorithms particularly for ball tracking [Wu et al., 2013]. We use two widely used metric for tracking evaluation.

Table 2.1: Center location error (CLE) result (CLE Threshold 10)

Video	Frames	Shift	ASLA	Mean-shift	Proposed
Seq 1	54	5	0.80	0.84	**0.87**
Seq 2	88	10	0.85	0.79	**0.83**
Seq 3	67	13	0.67	0.62	**1.00**
Seq 4	91	17	0.62	0.59	**0.81**
Seq 5	85	23	0.69	0.70	**0.78**
Seq 6	106	26	0.53	0.60	**0.74**
Seq 7	100	31	0.66	0.61	**0.75**
Seq 8	82	35	0.48	0.51	**0.79**
Seq 9	51	41	0.45	0.53	**0.76**

Precision Plot: Let us define the center location error (CLE) is the Euclidean distance between the center of the ball generated by algorithm and groundtruth. The precision plot is defined as the percentage of frames whose center location error (CLE) are smaller than a threshold [Wu et al., 2013]. Considering 5 pixels gap, the threshold is plotted along x axis in Fig. 2.6. However, the center location error only measures the pixel difference of the given location and the center respectively and does not reflect the size and scale of the

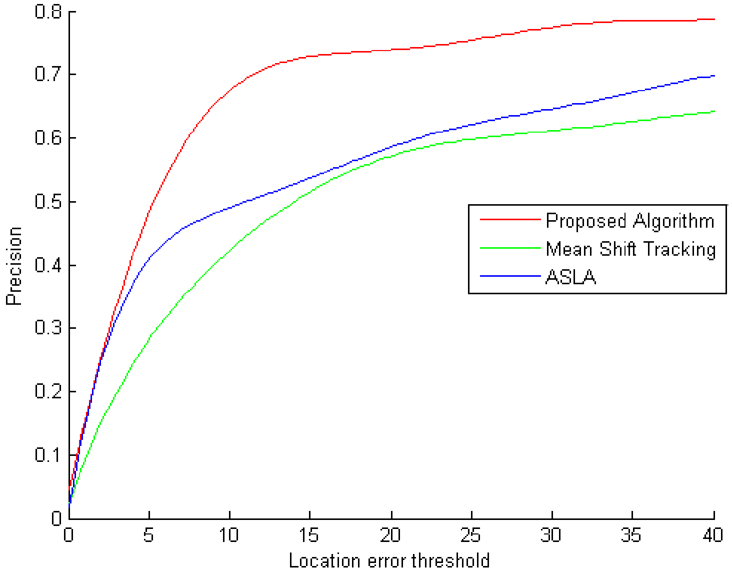


Figure 2.6: The x axis represents the CLE threshold used for comparison between the center of the ball in ground truth and our algorithm. The y axis represents the percentage of successful frames.

target object. In Fig 2.6 we have compared our method with color histogram based meanshift algorithm [Comaniciu et al., 2000]. Our method has shown better performance because we have used edge and shape of soccer ball with the color of soccer ball as our features. This method is 7.2% better than ASLA human tracking algorithm [Jia et al., 2012]. In Table 2.1 we take nine different situations from the data set. Here we have shown how accurately our algorithm is giving result than other algorithm. We consider maximum distance between the two circles, the groundtruth center and the proposed center. The considered distance is the pixel distance.

Success Plot: For completeness of the evaluation, the metric we use is the Area Under the Curve (AUC). Suppose we have a region R_t tracked by our tracking algorithm. We also have ground truth region R_g. By $|R_t|$ we denote the number of pixels in that region. Given R_t and R_g the overlap score is defined as

$$O_{score} = \frac{|R_g \bigcap R_t|}{|R_g \bigcup R_t|}.$$

The success rate of a tracker on a sequence is the percentage of frames whose overlap score O_{score} is larger than a given threshold. By varying the threshold from 0 to 1, one can generate the success plot, and the area under the curve (AUC) can be derived afterwards. In Fig. 2.7 we have shown our success rate. Our success rate is high as we have used weight assignment technique of particle filter and we resample coordinates based on the pointwise likelihood ratio. We also take shape and edge of the ball as our features along with the color feature which makes our algorithm more accurate than color histogram based mean shift algorithm and ASLA human tracking method.

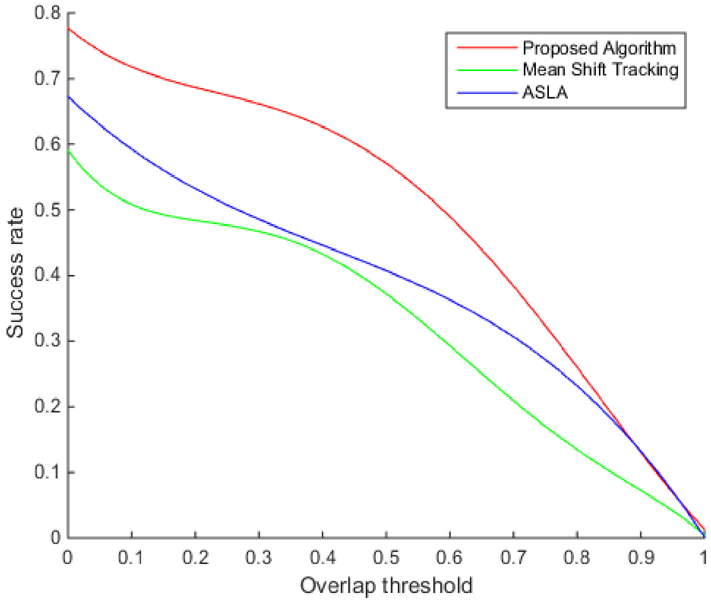


Figure 2.7: The x axis represents the overlap threshold between the ground truth and our algorithm generated region. The y axis represents the percentage of frames passes by the threshold plot in x axis.

2.3.4 Qualitative evaluation

We first conduct experiment to analyze the performance of our algorithm. The experiment is carried on tracking ball in different critical situation like sudden velocity change of ball, frequent change of ball's direction. We also make experimental environment critical by taking jersey color of one team white. Many players wear white shoes. It becomes difficult to track the ball when a player wearing white shoe takes the ball then pass it. We run our algorithm over 721 frames for all video sequences and .

We take frame 61 to 92 as our experiment frame. In these frames the football changes its orientation suddenly. In Fig. 2.8(a) we are showing our ground truth result where we manually label the ball with yellow circle. In Fig. 2.8(b) we run Meanshift algorithm based on color histogram [Comaniciu et al., 2000] which fails to track the ball during sudden direction changes. Basically, mean shift is a procedure for locating the maxima of a density function given discrete data sampled from that function. It is useful for detecting the modes of this density. So when the ball is in touch with a player with white shoe then it is unable to track the ball based on the mode of color based probability distribution function. In Fig.2.8(c) we run our football tracking algorithm which tracks the ball during sudden direction changes with high speed. This is because we give a weight on the shape of the ball.

In Fig. 2.9(a) we take frames 101 to 113 as our experiment frame. In these frames the football gets partial occlusion and changes orientation suddenly. In Fig. 2.9(b) we are showing our ground truth result where we manually label the ball with yellow circle. In Fig. 2.9(b) we run [Comaniciu et al., 2000] which fails to track the ball during sudden direction change. In Fig. 2.9(c) we run our football tracking algorithm which perfectly tracks the ball though partial occlusion occurs and the ball changes its direction with high speed suddenly. This

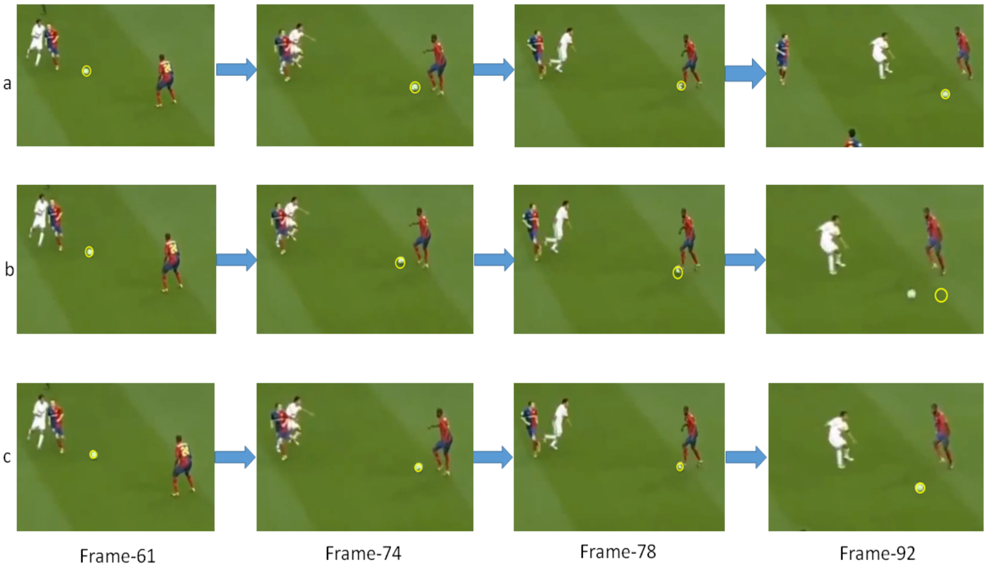


Figure 2.8: When the ball suddenly changes its direction with high speed at the last column frame 92: (a) Ground truth result, (b) Result of mean shift algorithm [Comaniciu et al., 2000] based on color histogram matching, (c) Result of our football tracking algorithm.

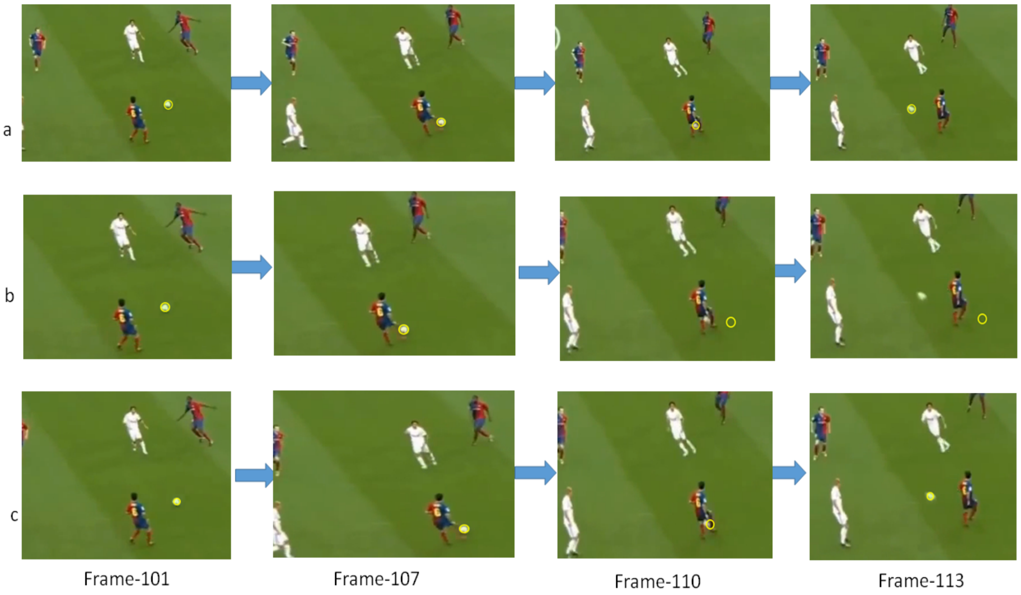


Figure 2.9: When the ball suddenly changes its direction and partial occlusion occurs, (a) Ground truth result, (b) Result of mean shift algorithm [Comaniciu et al., 2000], (c) Result of our football tracking algorithm.

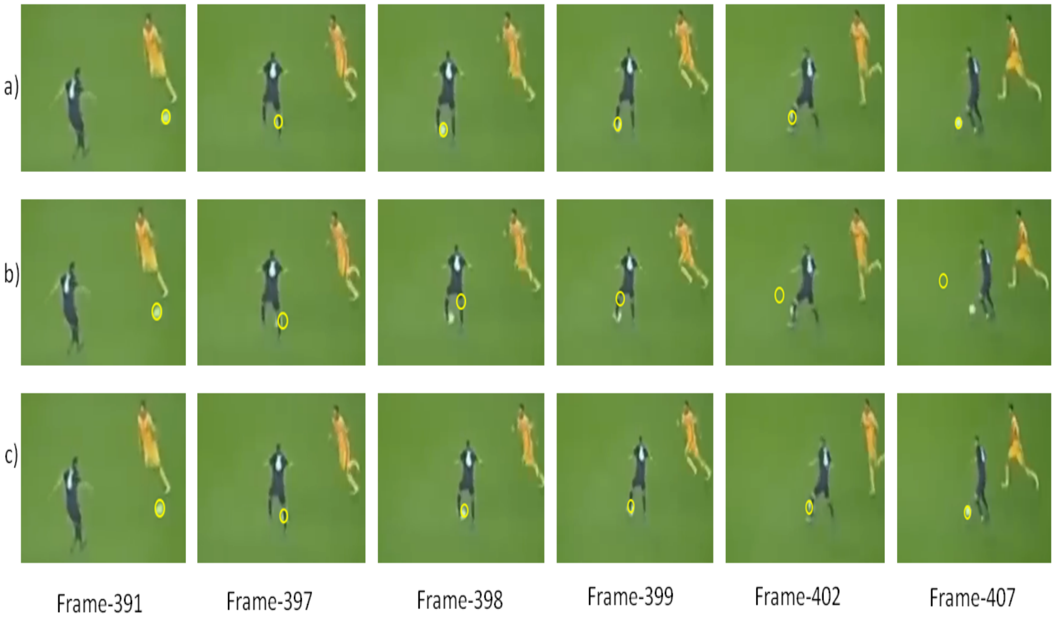


Figure 2.10: Qualitative evaluation when partial occlusion occurs at the column 3 and 4 respectively. (a) Ground truth result, (b) Result of mean shift algorithm [Comaniciu et al., 2000], (c) Result of our football tracking algorithm.

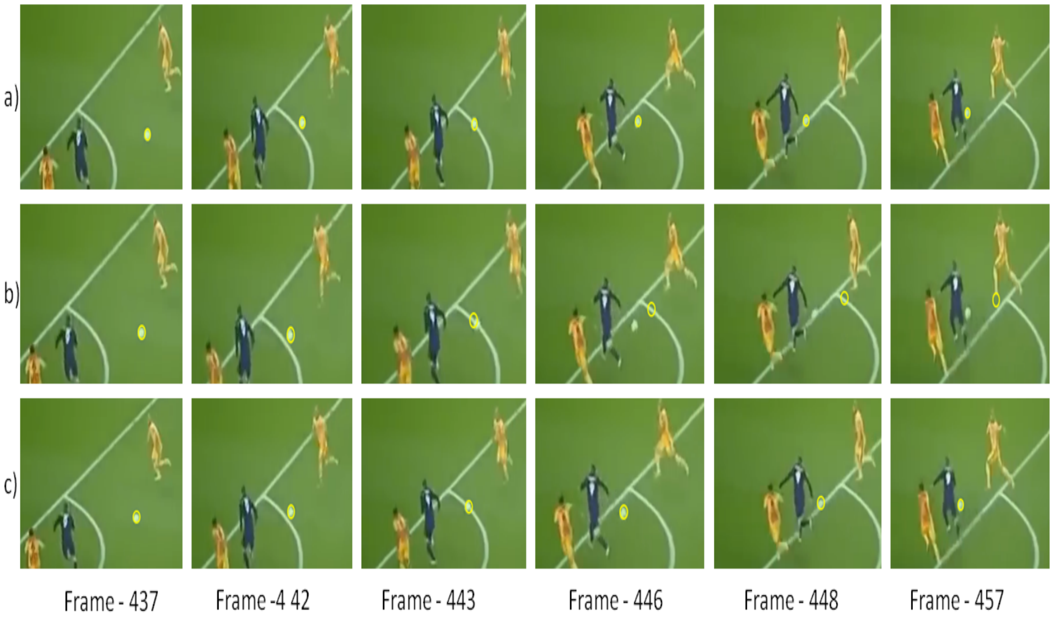


Figure 2.11: Qualitative evaluation when there are other small white objects like the white markings of the ground. (a) Ground truth result, (b) Result of mean shift algorithm based on color histogram matching, (c) Result of our football tracking algorithm.

is because we give weight on the gradient of the ball and we give attention to the size and shape of ball.

In Fig. 2.10(a) we consider frames 391 to 407 as our experiment frame of dataset [Dat, 2022a]. In these frames the football gets partial occlusion by player. In Fig. 2.10(b) we are showing our ground truth result where we manually label the ball with yellow circle. In Fig. 2.10(b) we run Meanshift algorithm based on color histogram [Comaniciu et al., 2000] which fails to track the ball when partial occlusion occurs. This is because for partial occlusion number of pixels on the ball is too low to track the ball based on color component. In Fig. 2.10(c) we run our football tracking algorithm which perfectly tracks the ball though partial occlusion occurs. This is because we give attention to the size and shape of ball and we assign weight based on size and shape of the ball.

In Fig. 2.11 we take frame 437 to 457 as our experiment frame of dataset [Dat, 2022a]. In these frames the football passes through the white line in the ground. In Fig. 2.11(a) we are showing our ground truth result where we manually label the ball with yellow circle. In Fig. 2.11(b) we run Meanshift algorithm based on color histogram [Comaniciu et al., 2000] which fails to track the ball passes through the white line in the ground. This is because the color of ball matches with the lines in the ground. In Fig. 2.11(c) we run our football tracking algorithm which perfectly tracks the ball though partial occlusion occurs. This is an example where our algorithm tracks the ball based on shape. Here gradient of edge pixel of ball and white line are almost same due to same background. Based on the shape of the ball we successfully track football when it passes through the white line.

2.4 Discussion

We propose a particle-filter based soccer ball tracking method. Along with the color and gradient information, we use shape measure to capture the multi-directional movement of the ball and track the ball when partial occlusion occurs due to the white lines of the ground and players. In order to develop the algorithm we propose a shape measure and incorporate with the color and gradient measures.

The next challenge is to derive the top-view visualization of the match that gives the exact position of the ball with respect to the whole ground. In the next chapter, we propose a top-view visualization approach that gives the exact location of the ball with respect to the whole ground.

Chapter 3

Top-View Visualization from Broadcast Video

3.1 Introduction

The whole paraphernalia of sports analytics is centered in the top-view visualization of the game. This visualization is a display of the game by an imaginary camera installed at the top of the center position of the ground where the game is being played. The top-view visualization provides the correct between-player distance which is distorted in the broadcasting video.

One way to create the top-view visualization is to equip the ground with many synchronized cameras. Thereafter, reconstruct the field and players in 3D using multi-view geometry techniques. Approaches of that spirit were previously proposed in the literature [Germann et al., 2012] and even commercialized as FreeD [Weblink, accessed March 2022]. The FreeD technology is used to enhance the quality of team sports like soccer, hockey, etc. However, the generality of the FreeD is limited due to multiple synchronized camera requirements. Another way to create the top-view visualization is to derive the map that registers players of the video frame into the abstract top-view model. To derive the registration map, a set of point correspondences between the video frame and the top-view model is needed. Finding the correspondence points requires locating, classifying, and tracking line intersections on the ground from the video [Sharma et al., 2018]. However, many broadcasting sports videos like soccer, hockey, etc. do not have such a precisely structured set of discriminatory features as the lines on the ground. To acquire such discriminatory features, static camera visualization of the game is helpful. The static camera visualization is a display of the game by an imaginary static camera installed at the sideline of the ground. Whatif we derive the top-view by transforming the broadcasting video into a static camera visualization as described in Fig. 3.1? Yes, this is the goal of the chapter as demonstrated in the video [1].

In the broadcasting video, the ground gets a relative motion due to camera movement. As a result, the correspondence points of the ground like the line intersections, corners, etc. disappears. Therefore, the intuitive idea is to transform the broadcasting video into a static camera visualization. Consequently, the ground becomes static, i.e. free from camera movement. Thus, the conundrum of the sudden disappearance of point correspondence is

[1] https://www.youtube.com/watch?v=tDSXqSjOMRU (Accessed Jan 2022)

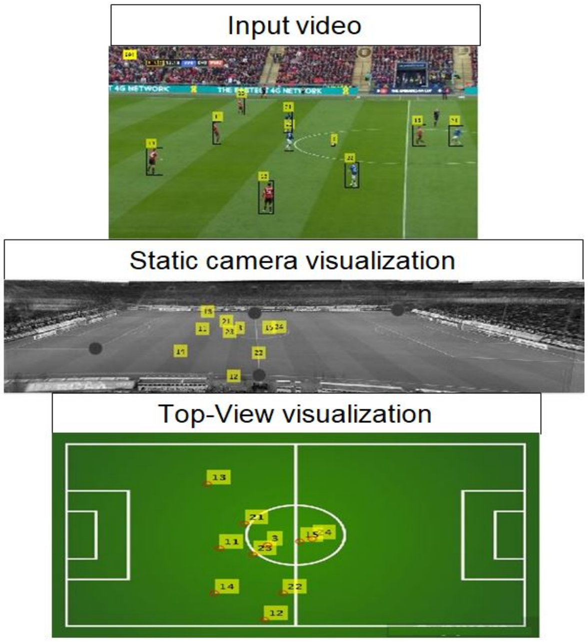


Figure 3.1: Let a broadcasting soccer video is given. First, we derive the single and static camera visualization from the broadcasting video. Thereafter, the top-view visualization is generated from the static camera visualization.

obliviated. In the work of [Sanyal, 2022], our contributions are as follows:

1. We propose a novel factor theory [Eckart and Young, 1936] inspired approach for the top-view visualization of the game from the broadcasting video. The factor-theory approach approximates a function as a product of two functions. Furthermore, we prove that the proposed approach of top-view visualization is better than the state-of-the-art approach with experimental validation.

2. A matrix completion paradigm [Candès and Recht, 2009] is proposed to reveal the global view of the ground, which is an important component of the proposed top-view model.

The remainder of the chapter is organized as follows. Section 3.2 presents the related work of the top-view visualization of the match. Section 3.3 presents the theory and real life implementation of the proposed top-view visualization approach. To engineer the proposed top-view visualization system, the global view of the ground is needed. In Section 3.4, first the paradigm to estimate the global-view of the ground is analyzed. Thereafter, the proposed top-view approach is experimentally analyzed and compared with the state-of-the-art. Section 3.6 summarizes this chapter and looks into future works.

3.2 Related work

Related work in the top-view visualization relies on either expensive hardware installation in the stadium or deriving the top view visualization from the available match video. Companies have proposed a hardware approach to top-view visualization problem [Weblink, accessed March 2022; Prozone, link accessed March 2022]. Advanced camera systems are installed in the gallery. However, the installation requires expensive equipment. Alternatively, companies such as Stathleates rely entirely on human workers for establishing the homography between the field and the model for every frame of the game [Stathlete, link accessed March 2022]. The computer vision based approaches derive the top-view visualization using the given video frame and the top-view model only. Next we discuss different compute-vision based approaches for top-view visualization that use the given video frame and the top-view model only.

Table 3.1: List of recent approaches for top-view visualization

Recent Works	Code name	Year	Comments
[Homayounfar et al., 2017]	DSM	2017	Proposed a CNN-based semantic detection to estimate vanishing points of a field in the image plane. Based on vanishing points, field registration is performed. The drawback of the approach is the difficulty of finding the vanishing points in real time due to unavailability of structured features of the ground.
[Sharma et al., 2018]	ATVR	2018	Proposed a method to register players by a nearest neighbour search over a synthetically generated dictionary of edge map and homography pairs. The drawback of the approach is the exploitation of the edge information from the line markings on the field which is not always available in every video frame.
[Chen and Little, 2019]	SCCS	2019	Proposed an automatic method for top-view visualization from a single image using synthetic data. The method detects field markings (e.g. the center line) present in a given video frame using generative adversarial network (GAN) model. The cons. of the approach is usage of synthetic data to generate camera pose engine which limits the approach for broadcasting video.
[Jiang et al., 2020]	OTLE	2020	Proposed a model that trains a deep network that regresses the registration error and then the images are registered by finding the registration parameters that minimize the registration error. The dependency on special features of the ground like the white lines may limit the efficacy of the model in a video.
[Sha et al., 2020]	EECC	2020	Proposed a method based on area-based segmentation, camera pose estimation and homography prediction via a spatial transform network (STN). The cons. of the approach is the semantic segmentation module which is trained with the groundtruth homography and not available in real time.

3.2.1 Field registration

Field registration is the method to generate reliable player coordinates inside the top-view. The top-view player coordinates are used in sports analytics [Sha et al., 2018], [Zhan et al., 2018], [Hobbs et al., 2018]. Fundamentally, in field registration method, the task is to find the homography that can map the 2D field from the video frame to a top-view model. Several methods like [Kim and Hong, 2000], [Farin et al., 2003], [Wang et al., 2006], [Watanabe et al., 2004] have been developed using geometric features of the ground such as lines and/or circles or arc. There are approaches based on Hough transforms [Yamada et al., 2002] and RANSAC [Carr et al., 2012]. Methods developed for field registration find correspondences between points and line segment [Carr et al., 2012], [Ghanem et al., 2012], [Gupta et al., 2011]. [Okuma et al., 2004] followed a frame-to-frame scheme where they calibrated each sequence using the initial homography and frame to frame matching. [Homayounfar et al., 2017] extended the works of [Hayet et al., 2004] and [Hayet and Piater, 2007] with CNN-based semantic detection to more accurately estimate vanishing points of a field in the image plane. The work of [Homayounfar et al., 2017] is extended by [Chen and Little, 2019]. [Chen and Little, 2019] used two generative adversarial networks (GANs) to extract the edges on a field, producing better reference image matching. The field registration is also facilitated by estimating the camera calibration.

3.2.2 Camera calibration

Recently an increasing number of approaches exploit the camera calibration to constrain the registration problem in broadcast video. Camera calibration is the process of estimating intrinsic and/or extrinsic parameters of the respective camera. In sports camera calibration, researchers assume the playing surface is flat. Consequently, the camera calibration is equivalent to estimating the homography from the playing surface to the image. Earlier works [Gupta et al., 2011] first manually annotate several reference images. Then, other images are related by finding correspondences from reference images. Fully-automatic methods are emerging because of a few requirement of user interactions [Wen et al., 2015]. [Chen and Little, 2019] estimated the base parameters and focal length of the camera to derive the top-view visualization. Initializing and refining the camera parameters or the homography is an approach discussed next.

3.2.3 Homography refinement

Previous approaches use methods like the Lucas-Kanade algorithm [Baker and Matthews, 2004] or Inexact Augmented Lagrangian Method (IALM) to refine the homography after its initialization from a matched reference frame with the top-view model. Some approaches are based on manual homography initialization for a given set of frames. Then compute the correspondence in subsequent frames. In order to avoid accumulated errors, the system needs to be reinitialized. Approaches of that spirit are used in [Dubrofsky and Woodham, 2008], [Hess and Fern, 2007]. A fundamental assumption of the mentioned approaches is that the transformation is small and local. [Carr et al., 2012], [Ghanem et al., 2012] formulated this problem as a non-linear optimization task, optimizing the homography between the given video frame and top-view model.

3.2.4 Camera pose estimators

Pose estimators are used to top-view visualization problem. Deep networks have been proposed to directly regress the degree of freedom of cameras pose matrix. Methods of that set up have been proposed in [Xiang et al., 2017], [Kendall et al., 2015] and [Mahendran et al., 2018]. However, these methods are dependent on their parameterization of the pose. So pose parameterizations can lead to bad performance, and are known to produce limited accuracy [Tekin et al., 2018]. [Jiang et al., 2020] use similar initial registration network of [Mahendran et al., 2018] following the same idea to sports field registrtion problem.

Table 3.2: List of symbols and abbreviation

Symbols	Meaning
F	Groundtruth top-view transformation
F^*	Arbitrary approximation of F by any state-of-the-art described in the Table 3.1
T	Embedding transformation from the given video frame to global view
H	Homography from the global-view to top-view model
G	Unknown global view of the ground
$(n_1 \times n_2)$	Dimension of unknown G
n	Minimum among n_1 and n_2
r	Rank we wish to recover of unknown G
$\mathcal{I}$	Given a set of images that covers the ground
m	The total number of available samples of the global view
X	Arbitrary image variable
$\|X\|_*$	The nuclear norm of the matrix X
S	Sample set containing location and value to the corresponding location
$P_S(G)$	The image with entries same as G at location given by S
O_κ	The operator applies a soft-thresholding rule to the singular values of the respective matrix effectively shrinking these towards zero
D_1	The benchmark dataset proposed in [Homayounfar et al., 2017]
D_2	The dataset proposed in this chapter
RE	Reconstruction error
IoU_{all}	Intersection over union over all over the ground
IoU_{part}	Intersection over union that is visible in the given image
M_1	Metric to evaluate the player registration performance
M_2	The metric to evaluate the quality of the top-view approximation map

3.2.5 Optimization with neural networks

Incorporating homography optimization into deep learning pipelines is a current topic of interest. The general idea is to learn an error function that predicts how well two images are aligned by homography. [Jiang et al., 2020] trains a deep network that regresses the registration error and then derive top-view visualization by finding the registration parameters.

All the discussed state-of-the-art methods of the Table 3.1 emulate the idea to derive the registration map directly from the video to the top view model. Therefore, the efficacy of the model depends on how well the models find out the correspondence points between the given video frame and top-view model. Finding point correspondence from the video becomes difficult due to camera movement and unavailability of high-level features of the ground. In contrast to the state-of-the-art methods, we transform the broadcasting video into static camera visualization. The static camera visualization contains all prominent point correspondence information of the ground. Therefore, with the help of static camera visualization, the top-view is derived. The proposed idea is inspired by the matrix factorization theory [Halko et al., 2011], [Martinsson and Voronin, 2016].

3.3 The proposed approach

We refer Table 3.2 for the list of symbols and abbreviations. The proposed idea for top-view visualization is described in the block diagram of Fig. 3.2. In Fig. 3.2(c) the detected players are transformed into the global view of the ground by the transformation T. Then the transformed players are registered into the top-view model by the transformation H in Fig. 3.2(d). Thus the top-view transformation F is approximated in factor form of H and T, i.e. $F \approx HT$. We claim that the proposed factorization approach is always better than directly

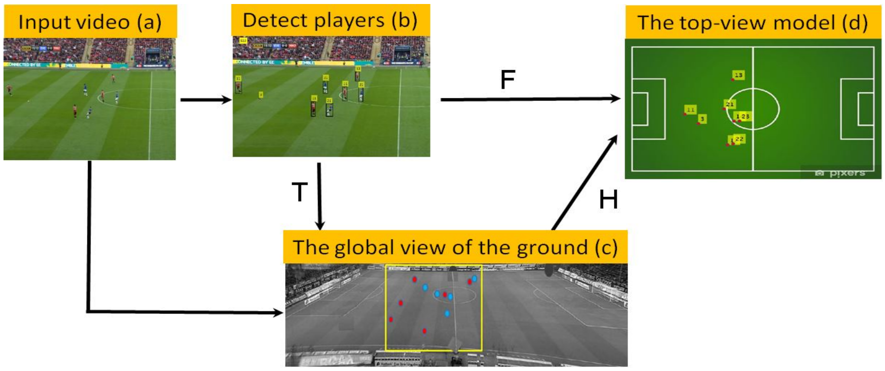


Figure 3.2: Block diagram of the proposed top-view visualization system. A soccer match video is given as input (a). Then frame by frame players are detected in the video frame (b) and the global view of the ground is generated (c). Thereafter, the given video frame is localized in the global view. Then the players of (b) are transformed in the global view with T (c). The fixed point correspondences like penalty box corners etc. are available between (c) and (d). Using the point correspondences, homography H is derived from (c) to (d). Thus the top-view transformation F is approximated as HT.

approximating F from Fig. 3.2(b) into Fig. 3.2(d). More precisely, for any arbitrary state-of-the-art approximation of F, we can always find a better approximation of F in HT form. We prove the claim in the following theorem 3.1 and experimentally validated in Section 3.4.4.6. The idea is inspired by matrix factorization theory [Halko et al., 2011].

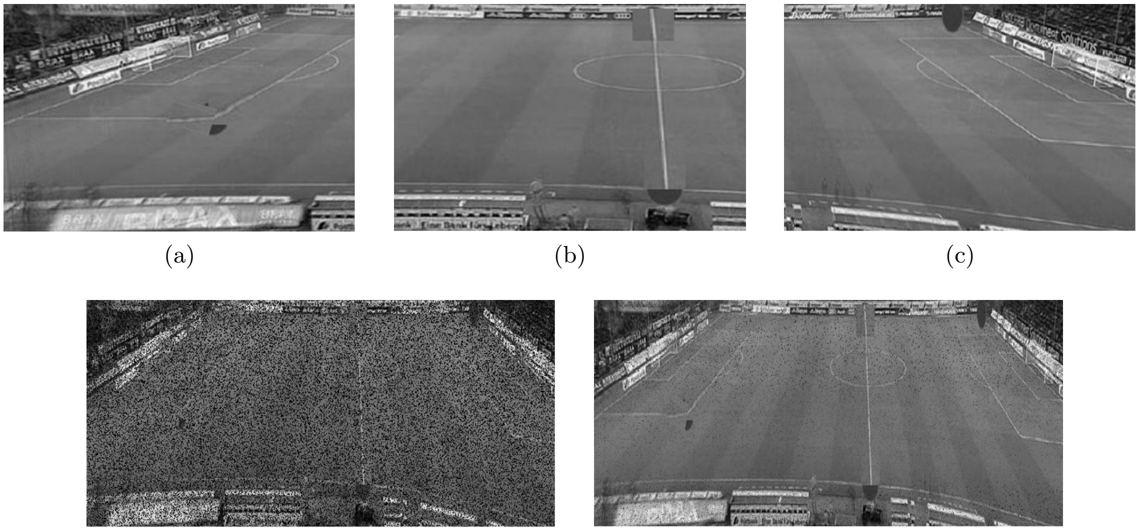
Theorem 3.1. *Let F^* be an arbitrary approximation of F directly derived by a model from (b) to (d) in Fig. 3.2. Then for any F^*, there exists a H and T such that (HT) is a better approximation of F than F^*.*

Proof outline: Let us assume that F is the groundtruth transformation that gives the top-view visualization. Given a video frame, the problem is to approximate F. Let the state-of-the-art method approximates F as F^* defined from (b) to (d) in Fig. 3.2. F^* registers players from the video frame to the top-view model with approximation error ϵ i.e. $\|F - F^*\| = \epsilon$, where $\|.\|$ is the Frobenius matrix norm. Given arbitrary F^*, we aim to construct T and H such that (HT) is a better approximation of F. Precisely, given arbitrary F^*, there exists T and H such that (HT) has approximation error less than ϵ. i.e. $\|F - (HT)\| < \epsilon$. The detail proof is given in appendix B.

The implementation of the proposed idea of theorem 3.1 is described in the block diagram of Fig. 3.2. To implement the proposed idea of Theorem 3.1, the entire ground or the global view of the ground is needed. Now the questions is: (1) how to find the global view of the ground? Furthermore, (2) How to derive T? and (3) How to derive H? The answers are expounded next.

3.3.1 Global view of the ground

The global view contains a complete image of the ground. A broadcasting soccer video may contain the global view of the ground as a single image. The first few minutes of any broadcast



(a) (b) (c)

(d) (e)

Figure 3.3: (a), (b) and (c) are the given images that covers the whole ground. (d) From the given images, the incomplete global view is generated by the algorithm 3.1. (e) The complete global view is generated by the algorithm 3.2.

video usually contain a global view of the ground without the presence of any player. On the other hand, a sequence of partial views of the ground may be used to create a global view.

However, that global view of the ground from a single image may not have prominent line information of the ground. The detail line information is only available when partial but clear images of the ground are available. Let us consider the example of a soccer match broadcasting video. The detailed images from all over the ground are captured by multiple cameras placed at different positions on the ground. So, the question is can we construct the global view of the ground from the images captured by different cameras as shown in Fig. 3.3? Towards the end, we introduce a new paradigm next.

Let the unknown global view image G has dimension $(n_1 \times n_2)$. We are given a set of L images say $\mathcal{I} = \{I_t\}_{t=1}^{L}$ that covers the ground. The images of $\mathcal{I}$ are captured by different cameras installed in the stadium. The $\mathcal{I}$ gives a few m pixel information of the unknown G, where $m < (n_1 \times n_2)$. Now the question is can we complete G from m known pixels?

3.3.1.1 The impetus for the solution

The issue with the global-view image completion is with a few known pixel values m compared to the total unknown $n_1 n_2$, we have many recovery solutions of G. For example, any given unknown location of G can be filled with any pixel value between 0 to 255. Therefore, it is apparently impossible to identify which of these recovery solutions is the correct one without some additional information. In [Candès and Recht, 2009], it is shown that the low-rank matrices (images) can be recovered exactly from a few number of known entries (pixels). Another key contribution is that they proved the matrix completion can be done by solving a convex optimization function. To state their results, let G be an $(n_1 \times n_2)$ unknown matrix. We have available m sampled entries $\{G_{ij} : (i, j) \in S\}$ where G_{ij} is the value at (i, j) location of the matrix G, S is a random m cardinality sampled subset of $\{1, 2, ..., n_1\} \times \{1, 2, ..., n_2\}$. Then [Candès and Recht, 2009] proves that matrices G of low rank r can be perfectly recovered by solving the optimization problem:

$$\underset{X}{argmin} \|X\|_*, \tag{3.1}$$

subject to $X_{ij} = G_{ij}$, $(i, j) \in S$, X_{ij} is the (i, j) location element of X provided that the number of samples $m \geq Cn^{6/5} r\text{log n}$ where $n = min(n_1, n_2)$, for some positive numerical constant C. $\|X\|_*$ is the nuclear norm of the matrix X, which is the sum of the singular values of X. However, to make the matrix completion problem tractable, [Candès and Recht, 2009] imposes some assumptions on the sample set containing the locations S and values at the corresponding locations. For our problem, the sampling set will be generated from $\mathcal{I}$. Next we will discuss the assumptions to design a sampling strategy.

3.3.1.2 Sampling assumptions

The assumptions in our context are as follows:

1. **Sample size:** To complete $(n_1 \times n_2)$ image G with $n_1 < n_2$ and rank less than or equal to r_1 needs atleast $4r_1(n_2 - r_1)$ pixels for unique solution when $r_1 \leq (n_2/2)$. Hence the sample set size $m \geq (4n_2r_1 - 4r_1^2)$ [Candès and Recht, 2009].

2. **Uniform sampling:** The sample set should be sampled uniformly at random from the given video frames. Every pixel of the given video frames has the probability $\frac{m}{n_1 n_2}$ to get included in the sample set. Uniform sampling guarantees that all rows and columns of G are sampled from the given video frames.

3. **Incoherence:** The incoherence assumption over the columns of an image ensures that any two columns have comparable amount of information. The assumption prevents any column becoming sparse and guarantees that the information is well-spread over the image G.

So, the assumptions of uniform sampling and incoherence jointly ensure that the sampled set captures the information of the unknown global view image G. Next we discuss the strategy to sample m pixel values from the given video frames.

3.3.2 The proposed sampling strategy

Given $\mathcal{I} = \{I_t\}_{t=1}^{L}$ of dimension $\{(n_1 \times p_t)\}_{t=1}^{L}$, where L is the total number of images that cover the ground. We have to sample m pixels from $\mathcal{I}$ to form the incomplete G. An example of $\mathcal{I}$ is shown in Fig. 3.3(a), 3.3(b) and 3.3(c). The sampling strategy should abide by the sampling assumptions. Abiding the sampling assumptions, the information in $\mathcal{I}$ should be uniformly spread. Therefore, each pixel is equally probable to be sampled. To select the samples uniformly, each I_t is converted into an array by concatenating the columns. Then the converted array I_t are concatenated to a value array of dimension $1 \times n_1(p_1 + ... + p_L)$.

Let $n_2 = p_1 + ... + p_L$. Now, we have to satisfy the uniformity and incoherence assumptions simultaneously during sampling process. So, first the value array is partitioned into n_2 bins each of size n_1. Simultaneously an index array is defined that keeps the location information

Algorithm 3.1 Sampling strategy

Input: $\mathcal{I} = \{I_t\}_{t=1}^{L}$. Create and partition the value array and index array into n_2 number of labeled bins of size n_1 as described in the section 3.3.2.

Iteration: For iteration 1 to n_2 perform:
(i) Randomly sample a bin label from $\{1, 2, ..., n_2\}$ without replacement (SRSWOR).
(ii) Randomly sample m/n_2 elements out of n_1 elements from the selected bin of index array and value array.

Output: The collection of all the sampled index array elements S with corresponding pixel values.

of each sample pixel. Each index array element consists of the information (array location, bin number). Now the sampling strategy is comprised of two simultaneous steps: (a) Simple random sample without replacement (SRSWOR) of a bin number and (b) random sampling of m/n_2 samples out of n_1 samples from the bin with sampled bin number. The step (a) ensures the uniformity of the sample selection all over the ground. Because the single array consists of all the ground information given by $\mathcal{I}$. The single array is partitioned by the labeled bins. SRSWOR of the bin numbers ensures the uniformity of the sample choices all over the ground. The step (b) ensures all the columns of the unknown G consist of equal number of sample pixels. Thus (a) and (b) take care of the sampling assumptions. The sampling process will stop after n_2 iterations as there are total n_2 labeled bins. The set of all the sampled index array elements is denoted as S. An example of the incomplete global view is shown in Fig. 3.3(d). The algorithm is presented in algorithm 3.1.

3.3.3 Some predicaments and the proposed solution

The proposed optimization function of Eq. 3.1 is capable to complete only the low rank matrix. In contrast, for our case the global view image is not low rank. So, only the optimization function of Eq. 3.1 can not estimate the global view image. Now the question is can we extend the Eq. 3.1 using the given sampled pixels, such that the incomplete G can be completed?

To recover the complete G from the given sampled pixels, the idea is to bring out the structure of the incomplete G and push towards the known sample pixel values at specified locations S. To elaborate the intuition mathematically, let X be an $(n_1 \times n_2)$ image variable. The Eq. 3.1 proposed by [Candès and Recht, 2009] reveals the structure of X by bringing out the magnitude of principal singular values [Candès and Recht, 2009]. Given the revealed structure of X, the idea is to push X towards the known pixels. Let P_S be function over the images that gives an image as output with the following property. If $(i, j) \in S$ then $[P_S(X)]_{ij} = x_{ij}$, where x_{ij} is the (i, j) location pixel of X and 0 otherwise. So, $P_S(G)$ gives the image with entries given by S. To push X towards the given pixels in S our strategy is as follows. For $(i, j) \in S$, we minimize the Euclidean distance between (i, j) pixel of G and X. Mathematically, we can write:

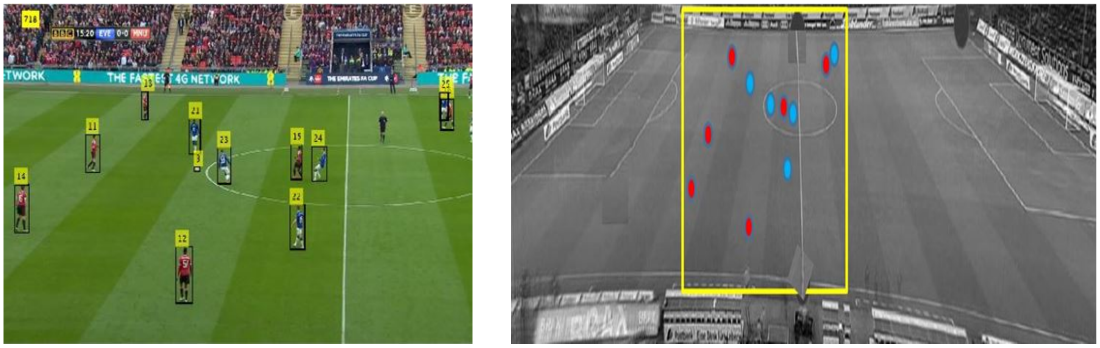
$$\underset{X}{argmin} \|P_S(X) - P_S(G)\|_F^2, \tag{3.2}$$

where $\|.\|_F$ is the Frobenius norm of a matrix. Combining Eq. 3.2 and Eq. 3.1, we can write the optimization function for the global view image estimation as follows:

$$\underset{X}{argmin}(\|P_S(X) - P_S(G)\|_F^2 + \kappa\|X\|_*), \tag{3.3}$$

where κ is regularization constant. The proposed Eq. 3.3 generalizes the capability of Eq. 3.1. More specifically, we claim that Eq. 3.3 is capable to complete the low rank images as can be obtained by solving Eq. 3.1. The claim is supported by the lemma 3.1.

Lemma 3.1. *Let X_κ^* be the solution to Eq. 3.3 and X_∞ be the minimum Frobenius norm solution to Eq. 3.1 defined as $X_\infty := \{\underset{X}{argmin}\|X\|_F^2 : X \text{is a solution of Eq. } 3.1\}$. Then* $\lim_{\kappa \to \infty} \|X_\kappa^* - X_\infty\|_F = 0$..



(a)

(b)

Figure 3.4: (a) In a given video frame, the players are detected. (b) The ground region of a given frame is localized inside the global view with yellow box as described in section 3.3.4. Thereafter, the detected players of the given frame are transformed by the transformation T, derived by the algorithm 3.3.

The proof is given in appendix B. In Eq. 3.3, κ plays an important role. If κ is very high then solution of Eq. 3.3 will prioritize to reveal the structure of the solution image only and then smooth the image with the help of given sample pixels. Consequently, we miss the intricate details of the ground like the white lines as described in the result section. Now, the question is can we keep intricate details of the ground while minimizing Eq. 3.3?

Short answer is yes (the ablation study described in section provides visual evidence in section 3.4.3.3). More elaborately, to keep the intricate details of the ground, $\kappa\|X\|_*$ should not be too high than the other term of Eq. 3.3 during the optimization process. Therefore the idea is to involve κ in minimization of all terms in Eq. 3.3. Keeping that in mind, we propose an iterative algorithm. Each iteration is comprised of two steps. The first step brings out the structure of the original image by eliminating small scale singular values using κ. The second step pushes the estimated image to the original image using the known pixels by implicit involvement of κ. Next, we describe the algorithm in detail.

The first step of the proposed algorithm is to reveal the structure of incomplete image using κ. The step is accomplished by the operator O_κ. We define the operator O_κ as: $O_\kappa(Y) = UO_\kappa(\Sigma)V^*$, where U and V^* are orthonormal matrices from the SVD of $Y = U\Sigma V^*$, $\Sigma = \text{diag}(\sigma_1, ..., \sigma_{r_2})$ contains r_2 singular values of Y, $O_\kappa(\Sigma) = \text{diag}(\{\sigma_i - \kappa\}_+)$. $\{\sigma_i - \kappa\}_+$

is the positive part of $\{\sigma_i - \kappa\}$. i.e. $\{\sigma_i - \kappa\}_+ = \max(0, \{\sigma_i - \kappa\})$. In words, O_κ applies a soft-thresholding rule to the singular values of Y effectively shrinking the singular values of low magnitude towards zero. Consequently, the principal singular values of Y are revealed.

The second step of the proposed algorithm is to push the output towards known entries. The pushing direction is the perturbation of the output from the known samples in S, $(P_S(G) - P_S(O_\kappa(Y)))$. Therefore, the second step of the proposed algorithm can be written as: $Y + \delta(P_S(G) - P_S(O_\kappa(Y)))$, where δ is the user given parameter. Starting with Y_0, the algorithm inductively defines:

$$\begin{cases} X_j = O_\kappa(Y_{j-1}), \\ Y_j = Y_{j-1} + \delta[P_S(G) - P_S(X_j)], \end{cases} \tag{3.4}$$

where $j = 1, 2, 3, \ldots$. We refer algorithm 3.2 for the global view estimation algorithm. The convergence of the Eq. 3.4 is supported by the following lemma 3.2. The convergence condition is

$$\frac{\|P_S(X_j) - P_S(G)\|_F}{\|P_S(G)\|_F} \leq \eta$$

and discussed in detail at the result section. Final output of algorithm 3.2 gives the global view of the ground.

Lemma 3.2. *For each $\kappa > 0$ and an $(n_1 \times n_2)$ image Y, algorithm 3.2 converges to: $\underset{X}{argmin}\|X - Y\|_F^2 + \kappa\|X\|_*$.*

In order to enhance the readability, we put the proof in the appendix B.

The algorithm 3.2 gives the global view of the ground. The next step is to detect players of the given video. Thereafter, the detected players are embeded inside the global view of the ground. Next we discuss how to detect players in the given video.

3.3.4 Player embedding into the global view

The tracking information of the ball can be obtained from the chapter 2. For the player detection, we use the method YOLO proposed in [Redmon and Farhadi, 2018]. The pretrained YOLO detector is available in the link [2]. Given the coordinates of the players of the

[2] https://pjreddie.com/darknet/yolo/

Algorithm 3.2 Global view estimation algorithm

Input: $P_S(G)$ and η.
Initialization: $Y_0 = \delta P_S(G)$, where δ is a user given parameter.
Iterative step: for $j = 1, 2, \ldots$, Stop if $\frac{\|P_S(X_j) - P_S(G)\|_F}{\|P_S(G)\|_F} \leq \eta$

$$\begin{cases} X_j = O_\kappa(Y_{j-1}), \\ Y_j = Y_{j-1} + \delta[P_S(G) - P_S(X_j)]. \end{cases}$$

Output: Estimated global view of the ground X_k.

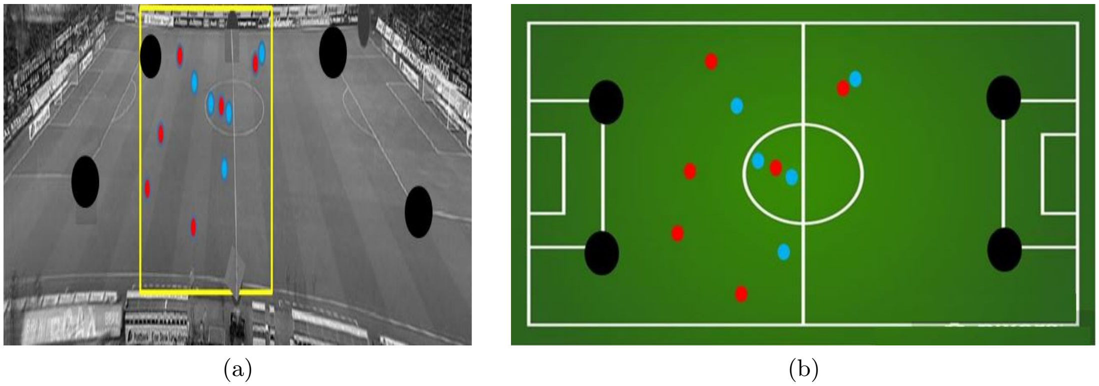


Figure 3.5: The corner points between the global view (a) and the top-view (b) are detected and matched. There after H is calculated and players of (a) are registered into (b).

video frame obtained by the tracking algorithm, the objective is to embed the players into the global view image as shown in Fig. 3.4. To define the embedding transformation, a set of matching points between the images of Fig. 3.4(a) and Fig. 3.4(b) are needed. However, the visual features of the ground are almost similar. As a result, some of the SIFT feature matching [Lowe, 2004] between the given images are false matching points. A possible way to obliterate the impediment of false matching is to find match points between a restricted region of the global view and the given frame. To find the restricted region inside the global view, the ground region revealed by the given frame should be found inside the global view as shown by the yellow box in the Fig. 3.5(a).

The embedding transformation is estimated in three steps: (a) Identify the ground region revealed by the given frame inside the global view. (b) Find matching points between given video frame and the identified region inside the global view and (c) iteratively estimate the transformation that maps maximum similar points. (a) To identify the given frame inside the global view, a YOLO model is trained with three different classes: The left side of the ground with the left half of the center circle, the middle of the ground containing the center line and circle and the right side of the ground containing the right half of the ground and the right half center circle. We use a total of 450 manually labeled images to train the YOLO

Algorithm 3.3 Embedding into global view (T calculation)

Input: Given video frame, classified region of G.

Points selection: SIFT feature match between the classified region of G and the video frame.

Initial step: Randomly select 4 matched points between given video frame and the classified region and calculate projective transformation $\hat{T}$.

Iteration: (i) Randomly select 4 matching point pairs transformed by $\hat{T}$.

(ii) Calculate $\hat{T}$ from the selected matching point pairs derived by iteration (i) and apply over all the SIFT matching points.

(iii) Find matched point pairs by $\hat{T}$ and continue step (i).

Output: $T = \hat{T}$ if $\hat{T}$ has more number of inlier points.

model. Then the trained model is used to classify the region revealed by the given frame inside the global view. The trained YOLO model outputs a bounding box inside the global view classifying the region revealed by the given frame. (b) Then a set of matching points are extracted between the given frame and the classified region of the global view using SIFT feature matching. (c) Using the matching point set the transformation T has been estimated iteratively. Here the RANSAC algorithm [Fischler and Bolles, 1981] is used in a modified way. First, randomly four matching points are selected. From the selected points a projective transformation is calculated. The transformation is iteratively calculated by all the available matching points as proposed by [Torr and Zisserman, 2000]. Thus we get the transformation T described in algorithm 3.3. T embeds all the player coordinates of the given frame inside the global view.

3.3.5 Player registration into the top-view model

The global view of the ground with players (Fig. 3.5(a)) and the top-view visualization model (Fig. 3.5(b)) are given. We have to construct the homography H that establishes the similarity between the global view and the top-view. In order to construct the H, atleast four point correspondences are needed. Then using the point correspondences the homography between the given images is defined. Corner points of the ground may be a good set of point correspondences between the global view (Fig. 3.5(a)) and the top-view model (Fig. 3.5(b)). So, the corner points of the ground are detected using Harris corner detector [Harris et al., 1988]. After the corner detection, the correspondence of the corner points is established using SIFT feature matching. Finally, with the help of correspondence point the H is derived [Hartley and Zisserman, 2003].

3.4 Experiments and results

Section 3.4.1 and 3.4.2 present the datasets used for our research. The empirical analysis of the global view estimation is presented in Section 3.4.3. Finally, the experimental analysis of the proposed factor-theory based top-view visualization approach is given in Section 3.4.4.

3.4.1 Dataset 1

We use the benchmark dataset for sports-field localization proposed by [Homayounfar et al., 2017]. The dataset is available in the link [3]. and denoted as D_1 in future discussion. The matches were held in 9 different stadiums. The images contain different perspectives and lighting conditions. There are 209 images collected from 10 different matches and 186 testing images from the other different matches. The dataset contains manually annotated correspondence points and manually computed homography between the collected images and a top-view model.

3.4.2 Dataset 2

To further demonstrate the power of matrix factorization approach, more elaborated datasets are needed. For example, to evaluate the quality of player registration inside top-view model,

[3] `https://nhoma.github.io/` (Accessed on 12.08. 2019)

player coordinates of the broadcasting video and the player correspondence inside the top-view model are necessary. However, the dataset proposed in [Homayounfar et al., 2017] does not have the required information. Therefore, a dataset has been developed from the dataset proposed in [Sanyal et al., 2016] and [Sanyal, 2021]. We have collected 5 soccer video sequences from La Liga and the English Premier League matches. Considering a gap of 3 frames, a total of 5000 frames of the soccer videos are annotated. The dataset exhibits a large range of lighting conditions. The games were recorded with multiple moving cameras. That results in varied camera motions and zooms. First, we label two teams as 1 and 2 respectively. We label bounding box of players of team 1 as 11, 12,...,111 and of team 2 as 21, 22,...,211. We label the bounding box containing the soccer ball as 3. In addition we derive the homography between the video frame and the abstract top view model by manual initialization of correspondence points. Thus a total of 5000 video frames are manually related with the abstract top-view by homography. We denote the proposed dataset as D_2 in next discussion.

3.4.3 The global view estimation

3.4.3.1 Parameters

1. **Sample size:** To reveal the $(n_1 \times n_2)$ dimensional global view image, we sample $(0.6 \times n_1 n_2)$ pixels from the $\mathcal{I}$. i.e. $m = (0.6 \times n_1 n_2)$.

2. **Stopping criteria for algorithm 3.2:** Inspired by Karush–Kuhn–Tucker (KKT condition) condition, we suggest the stopping criteria of the algorithm $\frac{\|P_S(X_k)-P_S(G)\|_F}{\|P_S(G)\|_F} \leq \eta$ where we take $\eta = 10^{-4}$ similar to the assumptions of [Cai et al., 2010].

3. **Choice of κ:** The rank of unknown G is not too low. So, we want $\kappa\|X\|_*$ slightly dominate the other term. The Frobenius norm of $\|P_S(X) - P_S(G)\|_F$ concentrates around $min(n_1, n_2) \times \sqrt{r}$ and the minimum nuclear norm concentrates around $min(n_1, n_2) \times r$ [Cai et al., 2010]. Therefore through out our experiments, we take $\kappa = 4min(n_1, n_2)$. Naturally, the proposed choice makes sure the value $\kappa\|X\|_*$ is greater than the other term for our experiments.

4. **Choice of δ:** The convergence of the algorithm 3.2 is guaranteed provided that $0 < \delta < 2$ [Cai et al., 2010]. A justification is as follows: Consider an $(n_1 \times n_2)$ dimensional image G. We assume that G satisfies the incoherence assumption. Then with higher probability over the choices of sample set, we can write [Candès and Recht, 2009]:

$$0 < \delta < 2\frac{\|G - X_k\|_F^2}{\|P_S(G) - P_S(X_k)\|_F^2} \tag{3.5}$$

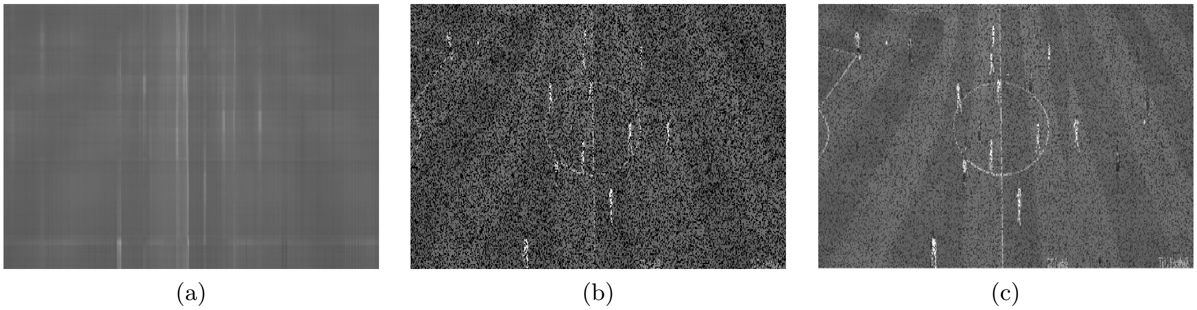
Since $\|P_S(G) - P_S(X_k)\|_F \leq \|G - X_j\|_F$, so we select $\delta < 2$. [Cai et al., 2010] shows that related matrix completion algorithms works the best when $\delta = 1.2 \times (\frac{m}{n_1 n_2})^{-1}$.

A study for the parameters κ and δ : In the Table 3.3, we present a study examining the role of the parameters κ and δ in the convergence of algorithm 3.2. In the Table 3.3, (400×1200) image of rank 300 is considered. Table 3.3 gives the number of iterations needed to achieve the proposed stopping criteria and the rank for different values of κ and δ. The

	$\delta = 0.8(\frac{m}{n_1 n_2})^{-1}$			$\delta = 1.2(\frac{m}{n_1 n_2})^{-1}$			$\delta = 1.6(\frac{m}{n_1 n_2})^{-1}$		
κ	# of iteration		rank	# of iteration		rank	# of iteration		rank
value	mean	sd	mean	mean	sd	mean	mean	sd	mean
$2min(n_1, n_2)$	632	203	350.4	714	1020	301.9	NA	NA	NA
$3min(n_1, n_2)$	417	5.6	300.0	180	3.8	300.0	1310	2194	300.0
$4min(n_1, n_2)$	443	4.2	300.0	195	2.4	298.5	266	593	300.0
$5min(n_1, n_2)$	471	3.2	300.0	315	3.8	300.0	145	5.2	300.0
$6min(n_1, n_2)$	507	7.4	300.0	412	2.7	300.0	310	6.4	300.0

Table 3.3: Mean and standard deviation (sd) over five runs of the number of iterations needed to achieve the stopping criteria of algorithm 3.2 for different values of the parameters δ and κ, together with the average rank of estimated image. The test example is image of dimension (400×1200) of rank 300. We report "NA" when none of the runs obeys the stopping criteria after $1,500$ iterations.

optimal δ is $1.2(\frac{m}{n_1 n_2})^{-1}$ as proposed by [Candès and Recht, 2009]. In the experiment of the Table 3.3 we choose the step size of δ and κ as $0.4(\frac{m}{n_1 n_2})^{-1}$ and $min(n_1, n_2)$ respectively. Table 3.3 implies that for each value of δ, there exists an optimal κ for which the algorithm 3.2 performs best. More elaborately, when κ is smaller than $4min(n_1, n_2)$, the number of iterations needed to achieve convergence is larger. When κ is close to the optimal value, the algorithm 3.2 exhibits a rapid convergence. However, there is little variability in the number of iterations needed to achieve convergence. If κ is too large, the algorithm 3.2 puts more importance on removing the intricate details of the estimated image at each iteration. As a result, the algorithm 3.2 exhibits slow convergence. Table 3.3 also indicates that the convergence of the algorithm 3.2 is dependent on the step size δ.



(a) (b) (c)

Figure 3.6: Ablation study of a (400×1200) image. (a) Minimization of $\|X\|_*$ removes the intricate details of the ground like the center lines and circle as noise. The RE score is 0.65. (b) Minimization of $\|P_S(X) - P_S(G)\|_F^2$ can not remove the noise from the ground. The RE score is 3.59×10^{-2}. (c) Minimization of $\|P_S(X) - P_S(G)\|_F^2 + \kappa \|X\|_*$ with the algorithm 3.2 preserves the intricate detail of the ground like center line and the circle as well as eliminates the noise. The RE score is 7.5×10^{-3}.

3.4.3.2 Reconstruction error

We define the relative error of the reconstruction as RE $= \|G - G_1\|_F / \|G_1\|_F$, where G and G_1 are the images generated by solving an optimization function and the groundtruth re-

Image attributes		Eq. (3.1)	Eq. (3.2)	Eq. (3.3)
Size ($n_1 \times n_2$)	Rank (r)	RE	RE	RE
50×150	10	1.64×10^{-3}	0.52	0.21
	15	1.58×10^{-3}	0.63	2.08×10^{-2}
	20	1.52×10^{-3}	0.72	1.89×10^{-2}
100×300	70	3.31×10^{-2}	0.63	6.24×10^{-3}
	80	2.78×10^{-1}	7.58×10^{-1}	3.08×10^{-2}
	90	7.52×10^{-1}	5.52×10^{-1}	5.62×10^{-2}
200×600	100	0.35	3.21×10^{-2}	1.32×10^{-3}
	150	0.46	2.58×10^{-2}	2.26×10^{-3}
	200	0.51	2.03×10^{-2}	3.21×10^{-3}
300×900	200	0.54	2.75×10^{-2}	3.73×10^{-3}
	250	0.58	4.65×10^{-2}	3.05×10^{-3}
	300	0.63	3.45×10^{-3}	4.62×10^{-4}
400×1200	300	0.62	3.73×10^{-3}	4.73×10^{-4}
	350	0.69	3.65×10^{-3}	6.65×10^{-4}
	400	0.74	2.62×10^{-3}	7.62×10^{-4}

Table 3.4: Ablation study for different size of images. The rank r is the rank of the images to be recovered. The columns 3, 4 and 5 provide the reconstruction error (RE) score of the estimated images by optimizing the Eq. 3.1, 3.2 and 3.3 respectively.

spectively. The RE evaluates the quality of the global view estimation.

3.4.3.3 Ablation study

We refer Fig. 3.6 for the ablation study of a (400×1200) dimensional image. The Fig. 3.6(a), Fig, 3.6(b) and Fig. 3.6(c) for the global view construction by minimizing $\|X\|_*$, $\|P_S(X) - P_S(G)\|_F^2$ and $\|P_S(X) - P_S(G)\|_F^2 + \kappa\|X\|_*$ respectively. The Fig. 3.6(c) shows that the proposed optimization function Eq. 3.3 works better than the other two validated by lowest RE score. For more elaborated analysis, we refer the Table 3.4. In the Table 3.4, $\underset{X}{\operatorname{argmin}}\|X\|_*$ performs well when the image has low rank. For example, the images of rank 10, 15 and 20 are estimated better by $\underset{X}{\operatorname{argmin}}\|X\|_*$ than $\underset{X}{\operatorname{argmin}}\|P_S(X) - P_S(G)\|_F^2$. However, the performance of $\underset{X}{\operatorname{argmin}}\|X\|_*$ deteriorates with the higher rank images. High rank images contain intricate details. For example the white lines of the ground in Fig. 3.6. $\underset{X}{\operatorname{argmin}}\|X\|_*$ removes the white lines by considering it as noise as shown in Fig. 3.6(a). In contrast the Eq. 3.2 performs better when the image rank is not too small as seen in the Table 3.4. For visual evidence we refer Fig. 3.6(b). This is because the number of correct sampled pixels increases with the resolution of the image. Minimization of Eq. 3.2 leads to produce images which have same values to the given pixel locations only. Consequently, the intricate details of Fig. 3.6(c) are more prominent than 3.6a. To appreciate the proposed Eq. 3.3 we refer the fifth or last column of the Table 3.4. The proposed Eq. 3.3 performs better than other two for higher rank images as seen in the second to fifth row of the Table 3.4. Even for the low rank images as presented in the first row of the Table 3.4, the proposed Eq. 3.3 is better than Eq. 3.2. For visual evidence we refer the Fig. 3.6(c). The intricate details of the ground are preserved in Fig. 3.6(c) as well as the noise is sufficiently removed. The possible reason is that the Eq. 3.3 combines Eq. 3.1 and 3.2. As a result we get enhanced performance.

Global view			Results		
Size ($n_1 \times n_2$)	Rank (r)	$\frac{m}{n_1 n_2}$	Steps	Time(s)	RE
250×600	150	0.42	177	5.7	1.64×10^{-4}
	200	0.55	165	8.2	1.58×10^{-4}
	250	0.67	158	10.3	1.52×10^{-4}
300×900	150	0.42	132	6.1	1.64×10^{-4}
	200	0.55	142	9.5	1.58×10^{-4}
	250	0.67	150	12.0	1.52×10^{-4}
400×1200	150	0.42	136	2.7	1.73×10^{-4}
	200	0.55	129	3.5	1.65×10^{-4}
	250	0.67	110	9.3	1.62×10^{-4}

Table 3.5: Simulation results for different size images. The rank r is the rank of the unknown global view, m/n_1n_2 is the fraction of observed entries. All the computational results on the right are averaged over five runs. For each test, the table shows the results of Algorithm 3.2.

3.4.3.4 Performance analysis of algorithm 3.2:

A numerical study of algorithm 3.2 is presented in the Table 3.5. We evaluate the computational times on Matlab 2019a for 64 bits system. We report the number of iterations the algo. 3.2 takes to reach the proposed error bound η, and the relative error (RE). To conduct the simulation, we use different resolutions of the global view image G as specified in the first column of the Table 3.5. The second column "rank" of the Table 3.5 reflects how much intricate information of the ground we want to preserve. The third column ($\frac{m}{n_1 n_2}$) of the Table 3.5 informs the fraction of the observed pixels. The column steps show how many iteration the algorithm takes to give the specific relative error. The results observed in the Table 3.5 are averaged over five runs. The first observation is that the proposed algorithm performs well in these experiments in terms of relative error (RE) score. In all of our experiments, the algorithm 3.2 takes fewer than 200 iterations to reach the proposed error bound of η. The second observation is the number of iteration decreases when the image resolution increases. The reason is that the algorithm brings out principal singular values due to the choice of high valued κ when image resolution is smaller. Consequently, to achieve the desired stopping criteria in low resolution images, the algorithm 3.2 takes a few more steps than higher resolution.

3.4.4 Top view performance analysis

To appreciate the performance of the proposed top-view model on the benchmark dataset D_1, we refer Fig. 3.7. In addition we introduce a new problem named player-registration problem. An example from our proposed dataset D_2 is given in Fig. 3.8.

3.4.4.1 Evaluation metrics

1. The evaluation metric defined by [Homayounfar et al., 2017] is denoted here as IoU_{all}. To define IoU_{all}, the $Area_{gt}$ and $Area_{alg}$ are the area covered inside the top-view model as per the groundtruth and the top-view algorithm respectively for a given video frame. Then the IoU_{all} can be computed as:

$$IoU_{all} = \frac{Area_{gt} \cap Area_{alg}}{Area_{gt} \cup Area_{alg}} \tag{3.6}$$

2. We denote the metric defined by [Sharma et al., 2018] as IoU_{part} metric. IoU_{part} is defined as the intersection over union of the ground part that is visible in the given image with the groundtruth. More specifically, the region of the ground defined by the white boundary lines is considered. The groundtruth localization is computed by corresponding the boundary lines visible in a given video frame and the top-view model. Let us consider the $Region_{gt}$ and $Region_{alg}$ are the region visible in a given video frame inside the top-view model as per the groundtruth and the top-view algorithm respectively. Then the IoU_{part} can be computed as:

$$IoU_{part} = \frac{Region_{gt} \cap Region_{alg}}{Region_{gt} \cup Region_{alg}} \tag{3.7}$$

3. However, the metrics IoU_{all} and IoU_{part} evaluate only the field localization quality of the algorithm. To evaluate how accurate the players are registered in the ground, we propose a metric M_1 as follows: Let $\{A_i\}_{i=1}^{Z}$ and $\{B_i\}_{i=1}^{Z}$ be the two dimensional top view coordinates of the Z number of players produced by algorithm and the groundtruth respectively. The evaluation metric M_1 is defined as:

$$M_1 = \frac{1}{Z}\sum_{i=1}^{Z} \frac{\|A_i - B_i\|_2}{\|B_i\|_2} \tag{3.8}$$

   M_1 computes the average relative error of the player position against the true position available in the groundtruth. M_1 evaluates how well an algorithm preserves the between-player distance by accurate player registration into the top-view model.

4. To assess the approximation quality of the top-view transformation, we define the metric M_2 as follows: Let us consider F and F_1 be the ground-truth and F_1 is the algorithm generated player registration transformation respectively. Then the metric M_2 is defined as:

$$M_2 = \frac{\|F - F_1\|}{\|F\|}, \tag{3.9}$$

   where $\|.\|$ is the Frobeneous matrix norm.

3.4.4.2 Competing methods

We compare with recent methods from all types of category described in the related works section 3.2. The state-of-the-art models of the Table 3.1 are chosen as competing methods.

3.4.4.3 Sports field localization performance

Identifying the field region revealed by a video frame inside the top-view is called sports field localization. The groundtruth homography localizes the ground region revealed by the

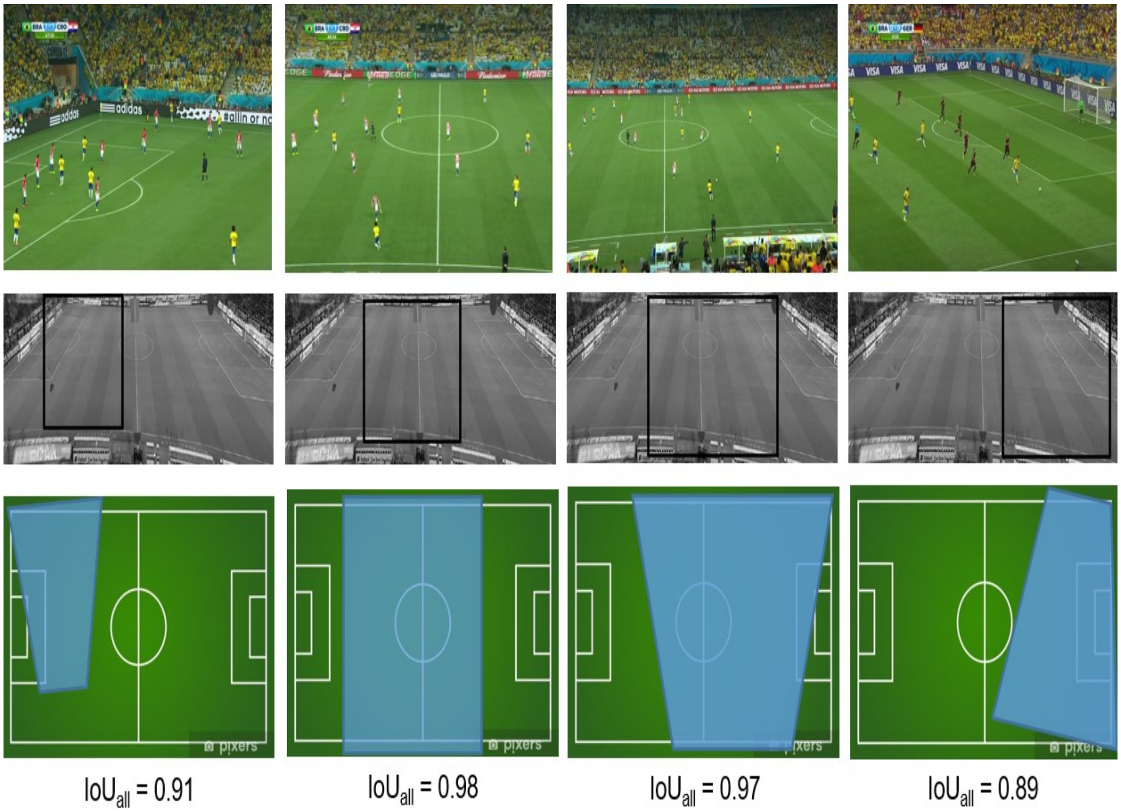


Figure 3.7: Field localization result on the benchmark dataset proposed by [Homayounfar et al., 2017]. First row: test images from the benchmark dataset. Second row: Given images are localized inside the global view. Third row: Localized region of the ground inside the top-view model. The quality of localization is described by the IoU_{all} score.

Dataset	DSM		OTLE		EECC		SCCS		Proposed	
Name	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median
D_1	83.0	—	89.8	92.9	88.3	92.1	89.4	93.8	**92.9**	**95.3**
D_2	70.0	72.2	74.8	75.2	73.7	76.2	75.4	76.8	**83.6**	**86.3**

Table 3.6: IoU_{all} score of the benchmark dataset [Homayounfar et al., 2017] and the proposed dataset. All recent works mentioned in the Table 3.1 DSM ([Homayounfar et al., 2017]), OTLE ([Jiang et al., 2020]), EECC ([Sha et al., 2020]) and SCCS (Chen and Little [2019]) are compared. The best performance is highlighted in bold.

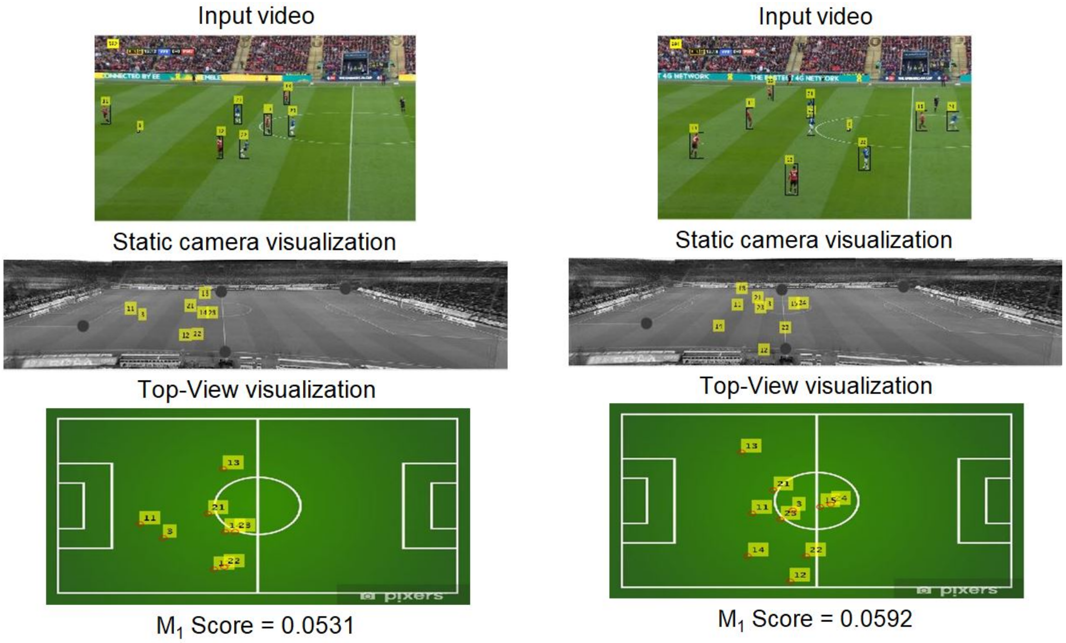
Dataset	ATVR		OTLE		EECC		SCCS		Proposed	
Name	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median
D_1	91.4	92.7	**95.1**	**96.7**	93.2	96.1	94.5	96.1	94.2	**96.7**
D_2	75.0	77.2	83.2	**84.5**	77.7	79.2	78.4	79.3	**83.6**	**84.5**

Table 3.7: IoU_{part} evaluation on the benchmark dataset [Homayounfar et al., 2017] and the proposed dataset. All recent works mentioned in the Table 3.1 ATVR ([Sharma et al., 2018]), OTLE ([Jiang et al., 2020]), EECC ([Sha et al., 2020]) and SCCS (Chen and Little [2019]) are compared. The best performance is highlighted in bold.

given video frame inside the top-view model. IoU_{all} accuracy reflects the performance on the field localization problem. The IoU_{all} accuracy using mean and median over the datasets D_1 and D_2 are presented in the Table 3.6. The proposed model improves the result of recent models by atleast 3% over D_1. The proposed approach also outperforms the recent models over D_2 by atleast 9%. The possible reason for the failure of [Homayounfar et al., 2017] is use of the vanishing point at every video frame. Atleast two sidelines are required to find the vanishing point which is missing in most of the frames of the broadcasting video. The possible reason for the failure of [Sha et al., 2020], [Chen and Little, 2019] [Jiang et al., 2020] is the dependency of the prominent and multiple boundary line. The boundary lines of the broadcasting video may not be extracted always due to insufficient discriminatory features. On the other hand, the proposed method relies on the localization of the given frame inside the constructed global view. The proposed approach is not dependent on a particular feature of the ground like the lines, texture etc. but on the overall information of the ground. This is the possible reason behind the success of the proposed approach.

3.4.4.4 Partial field localization performance

Identifying the ground part revealed in the given image is partial field localization. The problem is introduced in [Sharma et al., 2018]. IoU_{part} evaluates the performance on the partial field localization problem. We refer the Table 3.7 for the performance of the proposed algorithm against recent models. The proposed approach localizes the given image inside the global view. Therefore, the proposed method achieves close to the best result by [Jiang et al., 2020]. Our method performed significantly better than [Sharma et al., 2018], [Sha et al., 2020] and [Chen and Little, 2019], but approximately equal to [Jiang et al., 2020]. On the dataset D_1, our method suffers the insufficient training data, particularly for the field region detected inside the global view. Consequently, the mean accuracy is less than [Jiang et al., 2020].



(a) (b)

Figure 3.8: Player registration performance on the proposed dataset. For a given video, first the static camera visualization is derived. Thereafter, the players are registered into the top-view model. The quality of player registration into the top-view model is described by the M_1 score. A detail demo video is given in the supplementary material.

.

3.4.4.5 Player registration performance

The current work addresses the problem of player registration into the top-view model. The player registration problem is defined as to place the players of the broadcasting video frame into the top-view model while preserving the true between-player distance. The efficacy of a top-view model is rigorously examined when it is tested over the ability to preserve the isometry of players. To evaluate the player registration performance we introduce the M_1 metric. The M_1 metric calculates the relative error to register all players present in a given video frame. To perform the player registration the position of the players in the given video frame is needed. However the player annotation is unavailable in the benchmark dataset D_1. For this purpose, we develop the dataset D_2. The performance with respect to M_1 metric over D_2 is presented in the Table 3.8. We report the mean and median of the M_1 score over the total video frames of each video sequence in D_2. Less M_1 score implies better performance by the algorithm. Table 3.8 indicates the proposed model outperforms existing approaches in D_2. i.e. the column 'Proposed' has the least M_1 score in mean and median. The result implies that the proposed approach estimates the transformation F more accurately than others. A possible reason for the better estimation is the factor form approximation described in section 3.3. The proposed approach approximates F as composition of H and T, where the H and T are approximated separately.

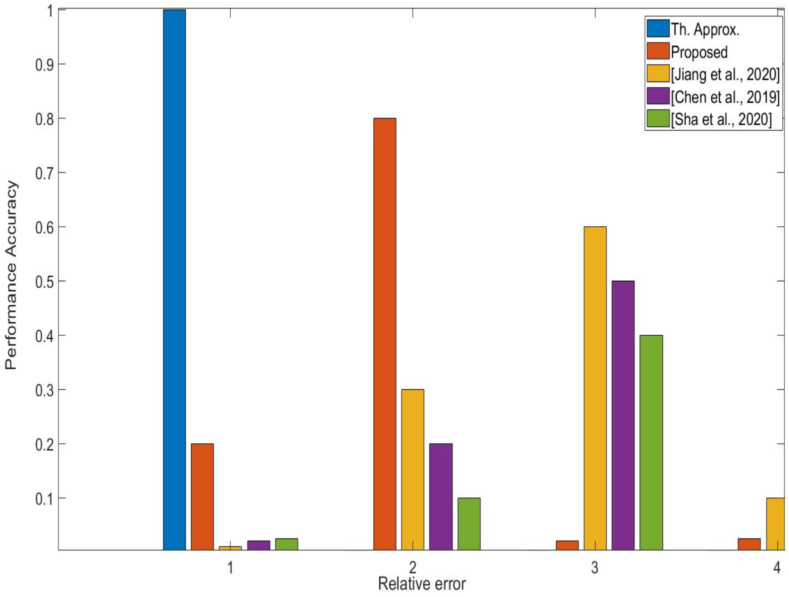


Figure 3.9: Approximation performance of the top-view map by different methods.The axes are described in section 3.4.4.6. "Th. Approx." is the theoretical approximation of the groundtruth top-view transformation of F mentioned in appendix B. 'Proposed' is the proposed top-view model in section 3.3. The state-of-the-art methods [Jiang et al., 2020], [Chen and Little, 2019] and [Sha et al., 2020] are considered for comparison.

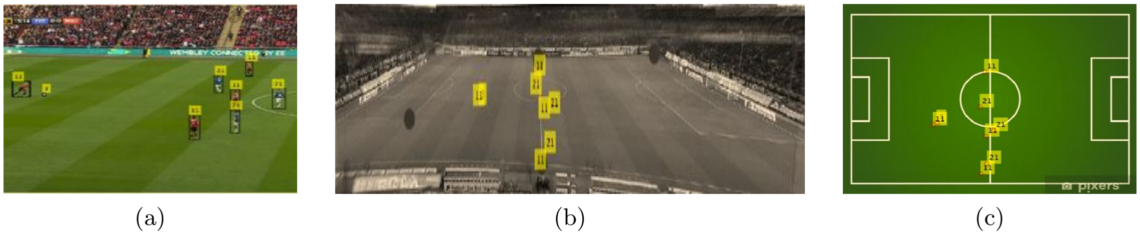
Dataset	ATVR		OTLE		EECC		SCCS		Proposed	
Name	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median
$S_1(735)$	0.2531	0.2571	0.1521	0.1528	0.3521	0.3527	0.1629	0.1633	**0.0591**	**0.0595**
$S_2(637)$	0.2030	0.2035	0.1923	0.1927	0.2521	0.2528	0.1527	0.1531	**0.0687**	**0.0692**
$S_3(1265)$	0.3214	0.3217	**0.1215**	**0.1219**	0.2972	0.2981	0.1925	0.1929	**0.1215**	**0.1219**
$S_4(1000)$	0.2006	0.2024	0.1185	0.1195	**0.1125**	**0.1131**	0.2521	0.2531	**0.1125**	0.1132
$S_5(1363)$	0.2126	0.2131	0.2138	0.2148	0.1567	0.1579	0.2417	0.2427	**0.1286**	**0.1291**

Table 3.8: M_1 score of player registration performance over all the video sequences of the dataset. The first column consists of the video sequence with total frames. All methods described in the table 3.1 ATVR ([Sharma et al., 2018]), OTLE ([Jiang et al., 2020]), EECC ([Sha et al., 2020]) and SCCS ([Chen and Little, 2019]) are compared. The table reflects the performance of the algorithms over 5000 video frames. The best performance is highlighted in bold. From the table we see that the proposed method outperforms other algorithms.

3.4.4.6 Approximation performance of the top-view transformation

In order to validate the theorem 3.1, we design the top-view registration map approximation experiment. First, let us consider the approximation error $\epsilon = 10^{-5}$. Thereafter, we derive the theoretical approximation of F as described in the appendix B. We compare all other registration transformation against the theoretical approximation of F. The state-of-the-art approach is directly approximate the top-view transformation from the given video frame. We choose the method [Jiang et al., 2020], [Chen and Little, 2019] and [Sha et al., 2020] as the representative of the state-of-the-art. Next, we describe the detail experiment.

We consider M_2 as evaluation metric for the registration map approximation. The M_2 metric score gives the relative error to approximate the top-view transformation F. Less M_2 metric score indicates better approximation of the F. The performance of the proposed approximation and the state-of-the-art is presented in a graphical form in Fig. 3.9. The graph presented in Fig. 3.9, we label the M_2 score intervals $[0, 10^{-5}]$, $[10^{-5}, 10^{-4}]$, $[10^{-4}, 10^{-3}]$ and $[10^{-3}, 10^{-2}]$ as 1, 2, 3, 4 respectively along the x axis. The proportion of frames corresponding to the labels are plotted along the y axis. In the graph of Fig. 3.9, The M_2 score of the theoretical approximation lies within the bin label 1. More elaborately, for all the 5000 frames, the theoretical approximation has relative error (M_2 score) in the interval $[0, 10^{-5}]$. This is because the M_2 score is the ratio of the approximation error and the Frobenious norm of the groundtruth top-view transformation. Here the approximation error of the theoretical transformation, ϵ is considered as 10^{-5}. Consequently, the M_2 score of the theoretical approximation becomes less and lies in the interval $[0, 10^{-5}]$ or label 1. Now we discuss the performance of the proposed model for the top-view visualization described in section 3.3. The M_2 score of the proposed top-view registration map lies within the intervals $[10^{-5}, 10^{-4}]$, $[10^{-4}, 10^{-3}]$. This is because the proposed top-view model is dependent on the localization of the given frame inside the global view. If the localization of the ground is not proper, then the approximation of the top-view map becomes erroneous. Consequently, the relative error or the M_2 score increases and the implemented top-view model gives more relative error than the theoretical approximation. However, the efficacy of the proposed top-view model is better than the state-of-the-arts. In Fig. 3.9 we see that the M_2 scores of the recent state-of-the-art models like [Jiang et al., 2020], [Chen and Little, 2019] and [Sha et al., 2020] lie mostly in the intervals $[10^{-4}, 10^{-3}]$ and $[10^{-3}, 10^{-2}]$ i.e. in label 3 and 4 respectively. A possible reason



(a) (b) (c)

Figure 3.10: An example of a failure case. (a) The discriminatory features of the ground are mostly unavailable. Therefore the players are not registered appropriately in the global view (b) and the top-view (c) respectively.

for the better performance of the proposed top-view model is the factor form approximation of the original top-view registration map.

3.5 Limitations

The proposed approach approximates the top-view registration map as a product of two linear transformation H and T. T embeds all players and the ball of the given frame into the global view. Therefore the embedded players in the global view are registered inside the top-view model by H. To derive the T, first the given video frame is localized inside the global view of the ground. Then SIFT points are searched between the given frame and the localized region inside the global view of the ground. Based on the matching points, T is calculated. If the given frame is not localized properly inside the global view of the ground, then the SIFT points may not be the desired one. Then the algorithm generated top-view transformation may not perform well as shown in the example Fig. 3.10.

3.6 Discussions

In this chapter, we have presented a factor theory based approach for the top-view visualization of the match from the broadcasting video. We theoretically prove that the proposed approach is better than the state-of-the-art, that estimates the top-view registration map directly from the video frame. The proposed top-view model derives the static camera visualization of the game from the broadcasting video with the help of global-view of the ground, estimated as a solution of convex optimization function.

We have shown thorough experiments that the proposed approach outperforms the state-of-the-art in the existing field localization problem on the benchmark dataset. In addition, to demonstrate the efficacy of the factor theory approach, we design rigorous experiments and necessary dataset towards the end. Our approach performs significantly better than the state-of-the-art in the proposed problem.

Thus in this chapter we develop the main component for visual analytic system. The top-view model provides the correct between-player distances. With the help of the between-player distance, a pass-prediction model is derived in the next chapter.

Chapter 4

Pass prediction from broadcast video

4.1 Introduction

Soccer is a team game that demands a well-designed strategy. The strategy in soccer includes increasing the number of valid passes in a game as opposed to number of incorrect passes. A valid pass is defined when a player of a team passes the ball to another player from the same team. The number of valid passes demonstrates the dominance of a team in the game. To prevent dominance, each team wants to limit the number of valid passes of the opposing team.

In order to limit the number of valid passes by the opposing team, one team is always trying to intercept the passes initiated by the other team. In other words, player of one team is trying to guess where the player of the opposing team who is trying to pass the ball before the pass is completed. The question is can a model be developed to guess the recipient of a pass initiated by a player. The development of this model is the objective of this chapter.

There are numerous challenges to predict the pass recipient in a soccer match directly from the broadcasting soccer video. Assuming that the players and ball are already identified in the video, prediction of pass recipient requires contextual knowledge. For example, we can predict with high confidence that a player tries to pass the ball to a teammate closer to the player but away from the opponent players. Further, a team might want to steer the game to a region where there is a better concentration of team mates. In this chapter our contributions are as follows:

1. Generation of the proximity model for pass prediction based on the positions of the opponent team players, and

2. Generation of pass region model for pass prediction that is influenced by the concentration of the players of the team who is in possession of the ball.

3. A fully annotated soccer dataset exclusively developed for data driven soccer analysis.

From the analytical perspective, the contributions are:

1. To propose a geometry based prior-free dependent model to measure proximity between players and opposition in a soccer game, and

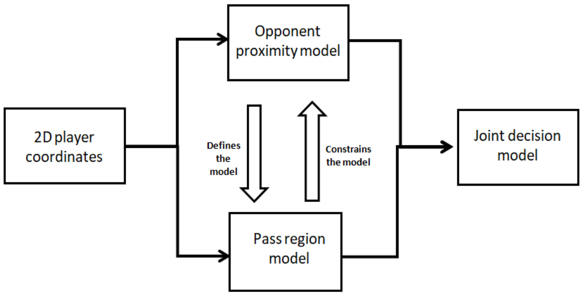


Figure 4.1: Block diagram of the proposed pass prediction framework. Player coordinates are given as input. We simultaneously develop opponent proximity model and pass region model for each pass aspirant. Finally, the decision is taken integrating two inter dependent models.

2. a theoretical framework to combine inter dependent models (proximity model and pass region model) to develop a prediction of pass.

4.2 Related works

Traditional works in computer vision analyzing sports videos have focused on either tracking key players [Ramanathan et al., 2016] or ball [Sanyal et al., 2016], [Maksai et al., 2016]. The survey [Gudmundsson and Horton, 2017] on team sports describes recent research efforts that involve spatio-temporal data as input. Another group of works assume the annotations of tracks of ball or players for analyzing game formations or skill level of individual players. For instance, [Rein and Memmert, 2016] analyzes soccer matches using player tracks and previous game information. [Bialkowski et al., 2016] discovers team formation in soccer match using player role representation instead of player identity. [Sarkar et al., 2019] calculates ball possession statistics using minimum flow network. [Wu et al., 2018] proposes a visual analytics system, which empowers users to track the spatio-temporal changes in team formation and understands how such changes occur.

Recently, research related to ball-passing becomes popular to analyze soccer strategy. The attribute pass has been used to understand how the collective performance may be optimized [Duch et al., 2010], [Lusher et al., 2010], [Grund, 2012] and [Fewell et al., 2012]. [Cintia et al., 2015] show how data-driven approach may present a big potential to accurately predict the team success. Indeed, the authors demonstrate that a view on football passing data has the potential of revealing hidden behaviours and patterns. [López Peña and Touchette, 2012] demonstrated that winning teams have higher fluidity of the ball passing dynamics between teammates. The insight helps to identify the key players involved during the offense and to describe their predominance of linkages within team positioning strategy at different levels of analysis [Gama et al., 2014], [Clemente et al., 2014]. Recently observation of pass and related events like ball possession becomes popular for soccer match analysis. [Link and Hoernig, 2017] has proposed models for detecting individual and team ball possession in soccer based on position data of ball and players. [Chawla et al., 2017b] propose a model to evaluate quality of a pass in soccer game. [Gonçalves et al., 2018] explore how passing networks and positioning variables can be linked to the match outcome in a soccer game. [Goes et al., 2019] has proposed a data driven model to quantify pass effectiveness by means of tracking data. The idea is to investigate how a pass disrupts the opposing defense. [Power et al., 2017] use a model based on an objective quantification of the risk and reward of a pass to assess the pass quality. [Rein et al., 2017] quantify pass effectiveness based on the change in space control in the final third of the field between the moment of the pass and the moment of the subsequent reception using Voronoi diagrams [Taki et al., 1996]. [Chawla et al., 2017a] proposes a system to automatically classify the pass on the field without linking it to goal scoring.

Closest to our work is the work of [Spearman et al., 2018] that uses physical concepts of interception and control time for pass prediction in soccer. However, their model is based on control time and velocity of players. Therefore the model fails due to abrupt velocity change of players. In contrast, we present a pass prediction model that works directly from visual inputs (coordinates of players). Naturally our model is robust to sudden velocity change of players. Another close to our work is [Felsen et al., 2017] that proposed a framework for forecasting future events in team sports videos directly from visual inputs. They predict pass from the visualization of top view of the game for waterpolo and basketball dataset. However, their framework does not incorporate contextual knowledge of the game. In contrast, our model introduces and exploits contextual knowledge like pass region to predict possible pass recipient. Next we present the pass prediction model.

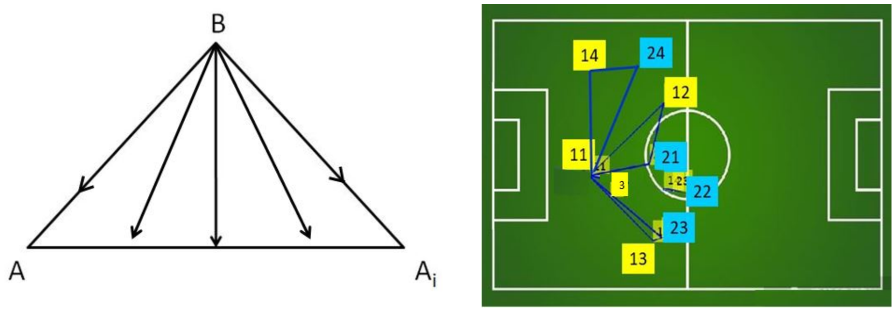
Table 4.1: List of symbols and abbreviations

Symbols	Meaning
OPM	Opponent proximity model.
PRM	Pass region model.
p	probability distribution derived from OPM.
q	probability distribution derived from PRM.
$PD(R)$	Player density of the region R
W_i	Normalized PD for ith region.
δ	Delta-Dirac measure.
R_P	Metric to evaluate robustness performance of pass prediction algorithm.
C_P	Metric to evaluate consistency performance of pass prediction algorithm.

4.3 Pass prediction model

The fundamental assumption considered to address the pass-prediction problem is that the ball will be passed to some players present in the respective video frame. Mathematically, let us assume a set of points in $\mathbb{R}^2$ moving on a 2D plane, is given. Given the contextual knowledge of positions of players from the two teams in a given video frame of a soccer match, can we predict the particular point (player) where the ball possessing player intends to pass the ball?

To answer this question, we propose two contextual information based models as described by soccer analyst [Stein et al., 2016]. One of the proposed models probabilistically quantifies the proximity of opponent team player(s) while another model estimates the quality of a pass region with maximum teammates. However, the proposed models are dependent on each other. Therefore, we propose a strategy that combines two dependent models. The block diagram of Fig. 4.1 introduces the proposed framework. Next we expound the framework in detail.



(a) (b)

Figure 4.2: (a) Player A has the ball. A can pass the ball to player A_i. Player B is the opponent team player. B can intercept A or A_i along the lines marked by the arrow heads. Thus the region-of-attack is defined by the triangle ABA_i. (b) Global scenario of opponent proximity model. Triangles are drawn considering two players of the same team and the nearest (smallest perpendicular distance to the pass line) opponent team player. For example, consider the pass line between players 11 and 12. Player 21 is nearest to the pass line among players 21, 22, 23 and 24. So, we consider the triangle joining the players 11, 12 and 21.

4.3.1 Proximity of opponent team player

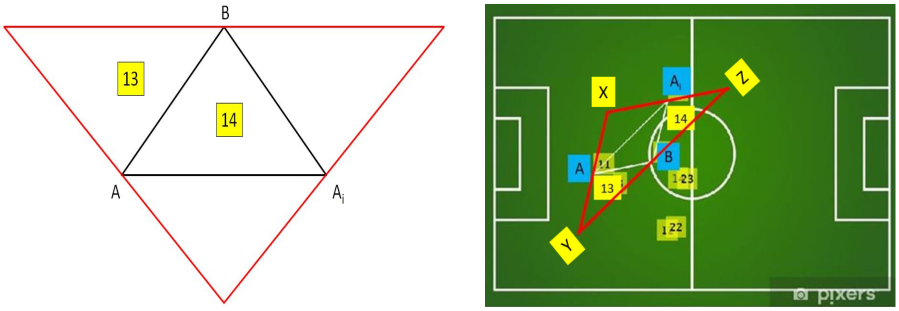
Let us consider the scenario in Fig. 4.2(a). The player A has the ball. A can pass the ball to player A_i (a representative teammate). Player B is the opponent team player. Let us consider a scenario when B lies on the pass line AA_i or around A_i. Then the probability of giving pass to A_i is very small as B can posses the ball easily. Another scenario is B is not on the pass line AA_i. Then B may attack A or A_i along any direction inside the triangular region ΔAA_iB as shown in Fig. 4.2(a). So the ΔAA_iB can be considered as the region-of-attack by the opponent player B. Assuming $area(\Delta AA_iB) > 0$, a model can be developed that approximates the probability of pass prediction given the proximity of opponent team player.

Opponent Proximity Model (OPM): We refer Fig. 4.2(a) for visualization of our model. Proximity of the player B to the player A and A_i can be measured by the area of the ΔAA_iB. Smaller area of the triangle indicates high proximity of the opponent team player. Consequently, ball passing probability of A to A_i will be less. So, we can write that the probability A_i gets the pass given the triangle ΔAA_iB ($p(A_i|\Delta AA_iB)$) is proportional to area of the triangle ΔAA_iB.

$$p(A_i|\Delta AA_iB) \propto area(\Delta AA_iB). \tag{4.1}$$

Now we generalize our proximity model for multiple opponents. First, we refer Fig. 4.2(b) for multiple opponent player scenario. In Fig. 4.2(b) player 11 can pass to players 12, 13, 14 and 15. We assume that opponent player whose perpendicular distance to the pass line is least, is most likely to tackle the pass line first. We form triangles considering pass lines as bases and the opponent player(s) as the other vertex of the triangle(s).

Thus, we get triangles in Fig. 4.2(b) for each possible teammates 12, 13 and 14. We assume that a ball possessing player is less likely to pass the ball to one of its far away teammates. According to formulation in (4.1), if a teammate is far from the ball possessing player, then the area of the triangle generated will be large. Consequently, the probability of giving pass to the far away teammate becomes high. In order to overcome this limitation, we introduce pass region model (PRM). Next we describe the PRM.



(a)

(b)

Figure 4.3: (a) Red triangle is drawn considering lines through the vertices A, A_i, B parallel to opposite sides A_iB, AB, A_iA respectively. (b) Global scenario of pass region model (PRM) for all possible pass players. The proposed pass region red triangle XYZ contains maximum number of team mates (13, 14 and A_i) of the ball possessing player (A).

4.3.2 Pass region model (PRM)

The model assigns a region where a ball possessing player is likely to give a pass. This region should have higher concentration of players from his team compared to the number of players from the opposing team. Therefore, the model needs to define a region in and around the triangle in Fig. 4.2(a).

We refer Fig. 4.3(a) for the visualization of the proposed pass region. At each vertices of ΔAA_iB, the parallel lines of the opposite sides are drawn. The lines intersect each other and form the red triangle XYZ of Fig. 4.3(b). The red triangle is denoted as the pass region $R(\Delta AA_iB)$. $R(\Delta AA_iB)$ has the possibility to contain a significant number of teammates when the area of ΔAA_iB is high enough. The observation is supported by the following lemma with the proof in Appendix C.

Lemma 4.1. *There are $P \geq 3$ points given in a plane. Say, no three points are collinear. Then all P points will lie in a triangle of area $\leq 4D$, where D is the largest area triangle among all the triangles formed by the P points.*

Player density (PD): Let there be k teammates of the player possessing the ball present inside the region R of the ground. Then the player density of the R denoted as $PD(R)$ is defined as $\frac{k}{area(R)}$.

Let the probability of A passing the ball to A_i given the pass region $R(\Delta AA_iB)$ is $q(A_i|R(\Delta AA_iB))$. By the definition of pass region, q is proportional to the player density inside $R(\Delta AA_iB)$.

$$q(A_i|R(\Delta AA_iB)) \propto PD(R(\Delta AA_iB)). \tag{4.2}$$

Let there be N teammates who can possibly receive the pass. For all the teammates, the proposed pass region is generated. The player density of each pass region is computed and normalized. Normalizing the player density corresponding to the pass region for N teammates, we get the sequence $\{W_i\}_{i=1}^{N}$ with $\sum_{i=1}^{N} W_i = 1$.

Now, the question is can we predict the possible pass recipients when a pass region is known? More specifically, how to approximate q with few known N estimates $\{W_i\}_{i=1}^{N}$ such that $\sum_{i=1}^{N} W_i = 1$.

The q can be approximated from a few estimation values $\{W_i\}_{i=1}^{N}$ with the help of Delta-Dirac measure δ [Arulampalam et al., 2002; Sanyal et al., 2016]. The δ is defined as equal to zero everywhere except at the estimated point of the approximated function $\int_{\mathbb{R}^2} \delta(x)dx = 1$. Using δ, the probability distribution of a player (represented as $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$) receiving the pass given the PRM, can be estimated as:

$$q(\mathbf{x}) \approx \sum_{i=1}^{N} W_i \delta\{\mathbf{x} - \mathbf{x}_i\}. \tag{4.3}$$

Next we discuss how the proposed pass region constrains our OPM model.

4.3.3 Constraining OPM with PRM

The proposed PRM is dependent on OPM by the construction of pass region. From (4.1), the proposed OPM depends only on the area of proximity triangle. Now, if a teammate is far away from the ball, then the current OPM gives higher probability of receiving a pass for the far away teammate. However, this is unlikely and limits the potential of OPM. To recover the limitation of the OPM, the proposed PRM plays a pivotal role. Now we discuss how PRM helps OPM to overcome the limitation.

The objective of the PRM is to establish the region of the ground with higher player density. The PRM exploits a pass region as shown in Fig. 4.3(a). In Fig. 4.3(a), if the distance between A and A_i increases, naturally $area(R(\Delta AA_iB))$ increases. Then, the probability to pass the ball to A_i decreases. We exploit this idea to constrain the decision due to OPM. Let us consider the line passing through B say $l(B)$ as the base for $R(\Delta AA_iB)$ and the height of $(R(\Delta AA_iB))$ as $H_T(B)$. If any of $l(B)$ or $H_T(B)$ increases then $(l(B) + H_T(B))$ increases. Accordingly, the OPM is penalized modifying (4.1) as:

$$p(A_i|\Delta AA_iB) \propto \frac{area(\Delta AA_iB)}{(l(B) + H_T(B))}. \tag{4.4}$$

Let there be N teammates who can possibly receive the pass. For N teammates, the proportional term of p in Eq. 4.4 is computed. Normalizing the proportional terms corresponding to N teammates, the sequence $\{\omega_i\}_{i=1}^{N}$ with $\sum_{i=1}^{N} \omega_i = 1$ is generated.

Now, the p can be approximated from a few estimated values $\{\omega_i\}_{i=1}^N$ as:

$$p(\mathbf{x}) \approx \sum_{i=1}^{N} \omega_i \delta\{\mathbf{x} - \mathbf{x}_i\}; \tag{4.5}$$

where δ is the Delta-Dirac measure.

4.3.4 Combination of two models

Here we combine the OPM and the PRM for joint decision of pass prediction. The state-of-the-art notion is to combine two independent models by multiplying the individual probability density function. In contrast to the state-of-the-art, we propose the joint decision scheme of two dependent models. The proposed OPM and PRM depend on each other. This is because the OPM defines PRM. So, the PRM is dependent on OPM by definition. In Eq. 4.4 the decision of OPM is influenced by the PRM. Thus OPM is dependent on PRM in decision making. Hence OPM and PRM are dependent on each other. Now the question is how to combine the decision of the two dependent models like OPM and PRM?

To combine two dependent models OPM and PRM, we refer Fig. 4.3(b). The player A has the ball. Consider the pass line AA_i. The player B is in the vicinity of the pass line AA_i. To extract the high player density region of the ground, PRM constructs $R(\Delta AA_iB)$. Inside $R(\Delta AA_iB)$, apart from A, available teammates are $A_i, 13, 14$. Now the question is how much the teammates are pressed by the opponents? The answer is given by the OPM. The proposed OPM measures the pressure on the players $A_i, 13, 14$ by the opponents. The next concern is to incorporate the benefit of the proposed pass region. The concern has been served by the PRM. Now, the joint probability of pass prediction can be derived by the multiplication of each distribution. Next, we elaborate the intuition mathematically.

Suppose the ball possessing player wants to pass the ball to a teammate $\mathbf{x}$. For $\mathbf{x}$ we generate a pass region. Inside the pass region, let k teammates $\{\mathbf{x}_j\}_{j=1}^k$ be available. How much the teammates get opponent pressure is given by the distribution p derived by OPM. Assuming the players are mutually exclusive, we can write the total pressure on the teammates of the pass region as $\sum_{j=1}^k p(\mathbf{x}_j)$. Now, the event described by OPM and PRM becomes mutually independent. So, the joint probability of getting pass for the player $\mathbf{x}$:

$$f(\mathbf{x}) = (\sum_{j=1}^{k} p(\mathbf{x}_j))q(\mathbf{x}); \tag{4.6}$$

We refer algorithm 4.1 for the pass prediction algorithm. Next, we discuss about the proposed dataset.

4.4 Soccer datasets

[Pettersen et al., 2014] have introduced a soccer video and player position dataset. The dataset consists of manually extracted ball positions. However, to evaluate the pass prediction algorithms, the correct accomplished pass information is needed. Therefore we introduce a video data set of soccer matches towards the end. The proposed data set contains ball and player coordinates, team information of each player, between-frame correspondence of players

Algorithm 4.1 Pass prediction algorithm

```
Input: Coordinates of the ball and players in the abstract top view visualization of current video frame.
Opponent proximity model (OPM): Compute $p$ for all $N$ teammates of ball possessing player given the nearest opponent using Eq. 4.4.
Pass region model (PRM): calculate the probability of creating good pass region for a possible pass accomplishment $q$ for all $N$ teammates by Eq. 4.2.
Joint decision: Calculate joint decision probability distribution $f$ using Eq. 4.6.
Output: Ball passed to the player id with highest score.
```

using unique id and details of valid pass for every frame. As mentioned earlier, a valid pass is defined when the ball goes from one player of a team to another player of the same team. Even though a pass may take several frames to complete, its origin and end are known. Given this, at every frame either the pass recipient or intended pass recipient is defined in the data set. Recent work of computer-vision based action spotting and event recognition in soccer [Giancola et al., 2018] and [Pappalardo et al., 2019] have proposed dataset. But, the dataset does not serve the pass prediction problem as it does not have player and ball coordinates. Pass prediction has been done based on raw data available from the video. A dataset of soccer video is proposed in [Spearman et al., 2018]. The dataset does not contain the id of pass recipient at every video frame. Considering all the above requirements, we have focused our efforts in designing the soccer dataset.

4.4.1 Dataset description

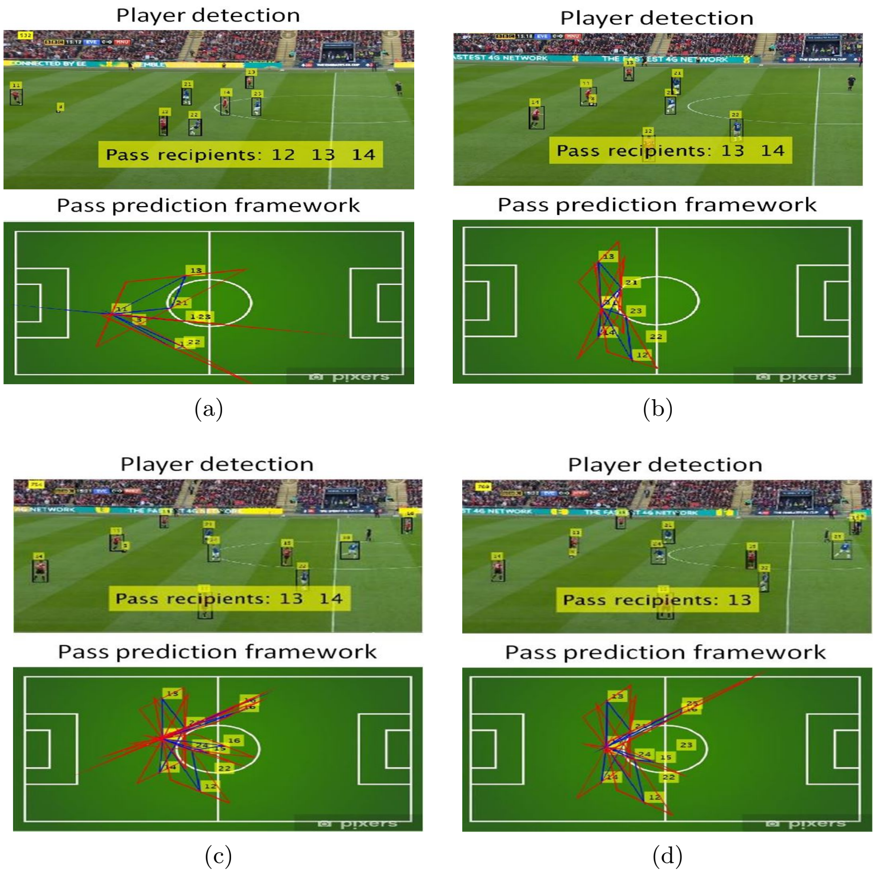
We have collected 5 soccer video sequences from La Liga and the English Premier League matches. The resolution of the videos is (1280 × 720) (720p) at 25 fps. The displacement of the players in consecutive frames is not significant. Therefore, to capture the significant between-frame player movement, we take 3 frames gap for annotation. Thus considering 3 frame gaps, we annotate a total of 5000 frames from the selected videos. The total annotated dataset is of 10 minutes. The dataset exhibits a large range of lighting conditions. The games were recorded with multiple moving cameras. Frames show zooming effect due to multiple cameras on the ground.

4.4.2 Player and ball annotations

First, we label two teams as 1 and 2 respectively. We label bounding box of players of team 1 as 11, 12,...,111 and of team 2 as 21, 22,...,211. We label the bounding box containing the soccer ball as 3. To identify and mark players, we manually run the video again and again and correctly identify correct player id. Thus we annotate a total of 5000 video frames with a variable number of players. The annotations of the players and ball will be helpful for the object tracking problem in sports video and data driven soccer strategy analysis.

4.4.3 Top view coordinates

We derive the homography between the video frame and the abstract top view model by manual initialization of four points of correspondence. Thus a total of 5000 video frames are



(a) (b) (c) (d)

Figure 4.4: Output of the proposed pass prediction system. (a), (b), (c) and (d) are the results extracted from the proposed dataset.

manually related with the abstract top-view by homography. The dataset details is discussed in chapter 3.

4.4.4 Ball-possessing player annotation

To identify ball possessing player, a ball possessing region is defined by automated algorithms. The ball possession of a player is defined in [Link and Hoernig, 2017] as the time duration for which the ball is in possession of a player. However, we find the neighbourhood radius plays the key role to define ball-possessing player. So, we assume 30 pixel radius circle from the feet of a player as ball possessing region of the player. Though it is difficult to identify the true player in ball possession due to overlap of ball-possessing region of players. To handle the overlap region issue, we manually annotate the player id in ball possession at frames where overlap occurs. Thus we annotate ball possessing player at every video frame. The information will be helpful for soccer strategy analysis.

4.4.5 Pass annotation

We annotate the true possible pass recipient id in all frames. Say a player id 11 has the ball in the 1st frame. He passes the ball to id 12 at 100th frame. We annotate true pass recipient id 12 for all 100 frames. Thus all 5000 frames are annotated. The information is necessary to evaluate data driven pass recipient prediction algorithm. In addition, the correct pass recipient is manually labelled for every video sequence. A total of 176 complete passes are manually annotated. The correct pass recipient annotation and the true pass labeling will be helpful for data driven pass variable analysis.

4.5 Experiments and results

We evaluate the performance of the proposed pass prediction algorithm. The proposed pass prediction algorithm is incorporated with computer vision based algorithms for object tracking and the top view visualization for real time application. Next, we discuss the implementation details of the pass prediction system.

4.5.1 Implementation details

Player detection model: In order to track the ball we use the method described in chapter 2. For the player detection, we use the method proposed in [Redmon and Farhadi, 2018]. The proposed method predicts bounding boxes representing players and a confidence score associated in deciding the class. The pre-trained YOLO detector is available in [1]. Darknet installation is required to run the pretrained YOLO model [2].

Top view visualization model: To obtain the top view from the broadcasting video, we apply the method proposed in chapter 3.

Parameters for the pass prediction algorithm 4.1: The proportional constant of Eq. 4.4 is evaluated as follows. Let there be total N pass aspirant players. For each pass aspirant, we calculate pass probability by Eq. 4.4. We calculate sum of all N pass probability values and use the inverted sum as proportional constant for Eq. 4.4. Similarly, we consider sum of outputs of Eq. 4.2 due to N total number of pass aspirants player. The inverted sum is used as proportional constant for Eq. 4.2.

4.5.2 Evaluation metrics

We propose two different metrics to evaluate the pass prediction algorithms. R_P metric: The pass prediction algorithm should predict the possible pass recipients at every frame. This is because the ball possessing player can pass the ball at any moment. Suppose the player 11 has the ball at the first frame. He passes the ball to 12 at the T th frame. Now the prediction algorithm is supposed to predict pass recipient 12 from first to Tth frame. Say the algorithm predicts correct pass recipient a total s_1 number of frames out of T frames. So, the R_P accuracy of the algorithm is (s_1/T). C_P metric: Suppose up to frame T, total u passes are played. Let the algorithm predicts u_1 correct passes out of total u passes. So, the C_P accuracy of the pass prediction algorithm up to the current frame T is (u_1/u).

[1] https://pjreddie.com/darknet/yolo/

[2] https://pjreddie.com/darknet/install/

Table 4.2: Robustness performance of the pass prediction algorithms on all video sequences of the dataset. The first column consists of the video sequence with total frames. The table reflects the R_P metric performance of the algorithms over 5000 video frames. From the table we see that the proposed algorithm outperforms other algorithms WWHN ([Felsen et al., 2017]) and PBMB ([Spearman et al., 2018]).

video (frames)	WWHN	PBMB	OPM	PRM	Proposed
S1 (735)	0.59	0.66	0.69	0.68	**0.75**
S2 (637)	0.56	0.50	0.67	0.70	**0.74**
S3 (1265)	0.51	0.58	0.69	0.71	**0.75**
S4 (1000)	0.49	0.40	0.69	0.71	**0.73**
S5 (1363)	0.51	0.42	0.65	0.67	**0.70**

4.5.3 Performance analysis and comparison

The pass prediction system must have three basic qualities. First, the system should predict possible pass recipients with high accuracy till the end of the given video. Second, the system should perform consistently at every given video frame. Third, the system should predict correct accomplished passes in a given video. To evaluate the mentioned qualities of the proposed pass-prediction system, we select recent pass prediction systems as discussed next.

4.5.3.1 Competing systems

We choose two recently published methods [Felsen et al., 2017] and [Spearman et al., 2018] for comparison. [Felsen et al., 2017] predicts pass using hand crafted feature derived from between player distance. The player coordinate from the video is extracted using pre-trained VGG-16 network [Simonyan and Zisserman, 2014] incorporated with Fast-RCNN [Girshick, 2015]. Another method [Spearman et al., 2018] proposes a time based pass prediction model. The model of [Spearman et al., 2018] exploits the time a player takes to reach the ball (time-to-intercept) and the time a player takes to control the ball (time-to-control). We have implemented the proposed pass prediction models of [Felsen et al., 2017] and [Spearman et al., 2018] and is denoted as WWHN and PBMP respectively in future discussion.

4.5.3.2 Robustness performance of the pass prediction systems

A pass prediction algorithm is robust if it performs well till the end of the video. i.e. high R_P accuracy at the end of the video. Here we discuss the RP of the pass-prediction systems.

To appreciate our algorithm, we compare our method against the paper [Felsen et al., 2017] and [Spearman et al., 2018]. In table 4.2 the performance of pass prediction algorithms is evaluated overall dataset videos (a total of 5000 video frames) and compared. The first column of the table 4.2 consists of video sequences with total frames. In table 4.2 the columns headed [Felsen et al., 2017], [Spearman et al., 2018] and 'Proposed' reflect the R_P score (robustness comparison). The results of table 4.2 show our method gives approx. 73% accuracy till the end of the video. The possible reason for the failure of [Felsen et al., 2017] is that they focused on team sports like basketball and water polo game video. The method does not consider the concept of possible pass region generation. On the other hand, our method exploits the region information to predict possible pass recipients. The proposed method of [Spearman et al., 2018] is dependent on time-based player movement. However, the time-based movement does not give valid results in the broadcasting video due to the variation of between frame change of player position.

4.5.3.3 Consistency performance of the pass prediction systems

A pass prediction algorithm is consistent if it performs well at every given time instance of the given video sequence. i.e. high R_P accuracy at the tth frame of the given video. We refer the graph in Fig. 4.5 to appreciate consistency of our proposed system over the dataset videos. The X and Y axes of the graph in Fig. 4.5 are the frame numbers and accuracy respectively. The graphs of Fig. 4.5, have two important characteristics. (i) The graph goes downward as the video frame number increases and (ii) the graphs in Fig. 4.5 are wavy. Next, we discuss (i) and (ii) in detail.

The reason of (i) is that the consistency performance is the ratio of the correct pass predicted frames say m_1 and t. Here, m_1 carries forward the past pass prediction error with the increment of t. So, the graph in Fig. 4.5 goes down with the increment of t. To appreciate our algorithm, we compare our method against the paper [Felsen et al., 2017] and [Spearman et al., 2018]. From the Fig. 4.5, we see the proposed method performs consistently as the 'proposed' curve goes uniformly against other models. The uniformity is reflected by the area under the curve (AUC) of Fig. 4.5. The possible reason for the inferior performance of [Felsen et al., 2017] is that the method does not consider the concept of possible pass region generation. In contrast, our method exploits the region information to predict possible pass recipients. So our pass prediction model performs more consistently against [Felsen et al., 2017]. The proposed method of [Spearman et al., 2018] is dependent on time-based player movement. However, the time-based movement does not give valid results in the broadcasting video due to the minimal between frame change. In contrast, our model exploits the geometry of the players at a given video frame. Hence, our model is not time dependent and robust against the minimal between frame change. The wavy nature of (ii) is caused by the tracking algorithm. When the tracking algorithm fails to track the players, then the graph goes down and successful tracking takes the graph upward. The possible reason for the wavy nature of [Felsen et al., 2017] is the pass prediction model and the tracking system developed with VGG 16 network. In contrast our system uses recent YOLO detector [Redmon and Farhadi, 2018]. This is the reason behind the proposed system curve is less wavy than [Felsen et al., 2017].

Table 4.3: PAA of pass prediction algorithms on all video sequences of the dataset. The first column consists of the video sequence with the total number of passes. The table reflects the C_P score of the algorithms over all 176 passes of the proposed dataset. From the table we see that the proposed algorithm outperforms other algorithms WWHN ([Felsen et al., 2017]) and PBMB ([Spearman et al., 2018]).

Video (passes)	WWHN	PBMB	OPM	PRM	Proposed
S1 (23)	0.77	0.70	0.80	0.84	**0.87**
S2 (22)	0.75	0.72	0.81	0.79	**0.83**
S3 (41)	0.78	0.75	0.82	0.80	**0.85**
S4 (35)	0.69	0.70	0.75	0.77	**0.79**
S5 (55)	0.70	0.71	0.76	0.79	**0.82**

4.5.3.4 Pass accomplishment accuracy (PAA)

A pass prediction algorithm or system can predict an accomplished pass efficiently if it has high C_P accuracy. Here we discuss and compare the pass accomplishment accuracy of the pass prediction models.

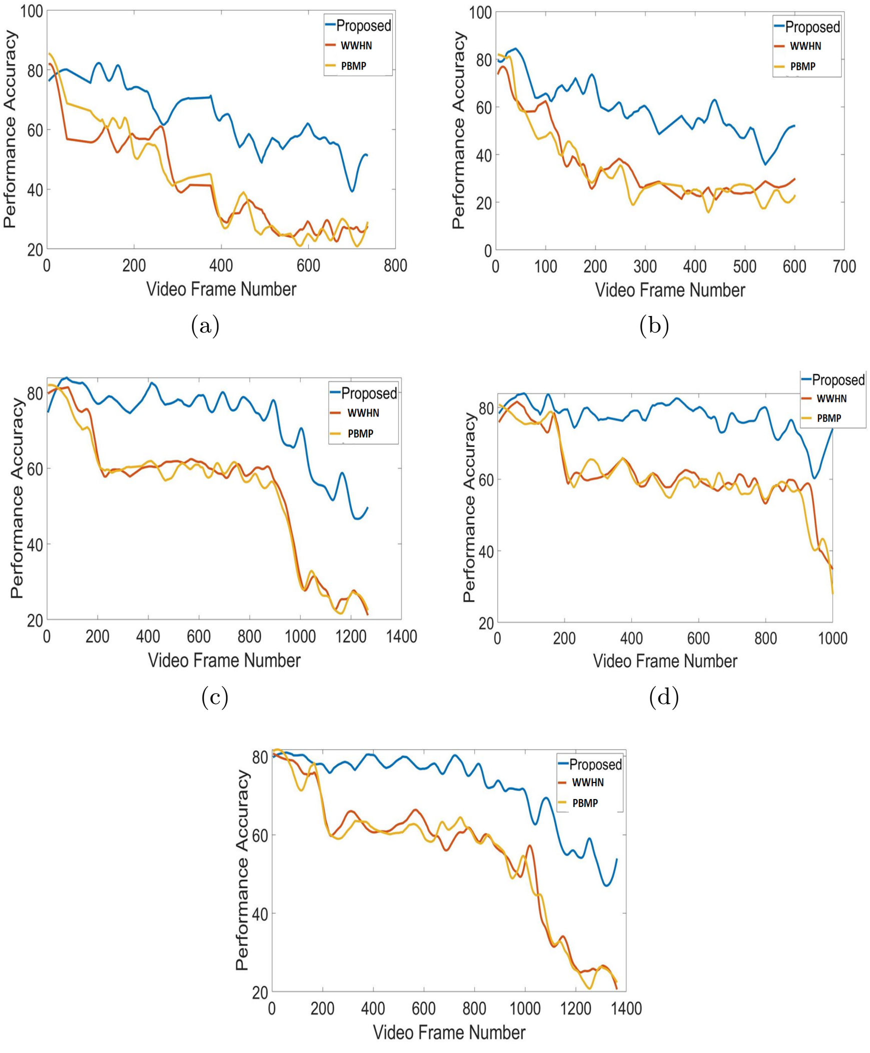
To appreciate the efficacy of the complete pass prediction ability of the proposed algorithm, we refer table 4.3. The proposed algorithm outperforms the pass prediction model of [Felsen et al., 2017] and [Spearman et al., 2018] in terms of C_P score. In table 4.3 the performance of pass prediction algorithms is evaluated over all dataset videos (a total of 176 complete passes) and compared. The first column of the table 4.3 consists of video sequences with total complete passes. In table 4.3 the columns headed [Felsen et al., 2017], [Spearman et al., 2018] and the proposed reflect the C_P score (pass accomplishment score). The results show our method gives on an average 84% accuracy over all videos. The average C_P accuracy of the pass prediction models of [Felsen et al., 2017] and [Spearman et al., 2018] are 73% and 71% respectively. Clearly, the proposed model gives 11% and 13% improvement over the state-of-the-art. The reason is that the proposed algorithm is more robust and consistent than other pass prediction algorithms. So, the proposed algorithm unveils the true pass recipient at the beginning of the pass more efficiently against other algorithms proposed in [Felsen et al., 2017] and [Spearman et al., 2018].

4.5.3.5 Ablation study

To appreciate the robustness and pass accomplishment of our combined model approach, we refer the last three columns of tables 4.2 and 4.3 respectively. From the result it is clear that the combined model outperforms individual decision models OPM and PRM. OPM can not show the robustness as there is no constrains over the area of the triangle it used. So, far away teammates get higher probability of getting pass than nearby teammates. The limitation of PRM is that it is inversely proportional to the region covered. So, in spite of covering more players, decision of PRM gets penalized if more ground area is covered. In contrast to the individual OPM and PRM, the combined model takes care of the limitation of both OPM and PRM. In order to overcome the limitation, OPM and PRM depend on each other. The dependency is exploited in the joint decision model of the proposed approach. As a result, the efficacy is reflected in the column 'proposed' of the tables 4.2 and 4.3 respectively.

4.6 Discussions

In this chaapter we present a pass prediction framework as well as an algorithmic approach to estimate pass prediction probability from a broadcast video. Two dependent models are combined to arrive at a decision. The proposed model is tested on a benchmark soccer dataset developed for pass prediction. The dataset can be used to validate ball possession statistics during a game and tracking challenge. We intend to work on ball possession statistics. Future work should also include pass networking and pass quality evaluation.



(a)

(b)

(c)

(d)

(e)

Figure 4.5: (a) - (e) are the Consistency performance of the pass prediction systems over all the dataset videos. We compare results with methods [Felsen et al., 2017] and [*Spearmanet al.*, 2018]. The results show that the proposed approach outperforms other methods with high AUC.

Chapter 5

Conclusions and Future Directions

5.1 Conclusions

We derive a pass-prediction model for a soccer match directly from the broadcasting video. The proposed pass-prediction model is comprised of several components. In the beginning, we propose a ball tracking model. Next, we derive a top-view visualization of the match from the broadcasting video. Finally, we propose a pass prediction model to predict possible pass recipients. From our research, presented in this thesis, we arrive at the following conclusions:

- State-of-the-art tracking algorithms can track the players in a broadcasting video very well. But ball tracking becomes difficult due to smaller size and high speed movement of the ball. Therefore the standard tracking algorithms often fails to track the ball as the algorithm requires incorporation of special components in the tracking algorithm.
- In order to derive a pass prediction system, the between-player distances are necessary. As the broadcasting video is a 2D projection of a 3D event, the correct between-player distances can not be realized from the broadcasting video. Therefore, to acquire the correct between-player distance, we propose a top-view visualization model inspired by the factor-theory. Furthermore we mathematically prove that the proposed approach for the top-view visualization is better than the state-of-the-art. To validate the proposed approach, we create a fully annotated football dataset.
- Given the player co-ordinates and the organization of players, the possible pass recipients can be predicted. The contextual information of a game scenario is derived from the organization of players. Towards this end, we propose models to derive the contextual information. Finally with the help of contextual information we predict the possible pass recipients by integrating the proposed models.

5.2 Challenges and future direction

While working on deriving a pass prediction system, it was possible to visualize challenges being opened up due to our research. These challenges may generate quite a few direction where research could be pursued. Some challenges that need attentions are as follows:

- In this thesis we propose a top-view visualization approach inspired by factor theory and prove the efficacy of the proposed approach over the state-of-the-art. The proposed approach approximates the top-view map as a product of two transformations. Therefore approximating the transformations in a more efficient way could be a future direction of research.

- The proposed pass prediction system uses a top-view visualization model which is 2D. The system can be enhanced by transforming it into a 3D visualization with the help of some additional hardware like hololens [Garon et al., 2016]. Therefore deriving a 3D pass recommendation system from the broadcasting video could be an obvious direction of research.

- Recommending the possible pass network is an important problem of strategy recommendation. With the help of possible pass recipients, several pass networks can be predicted. Hence in future we would focus on predicting possible pass networks from the broadcasting video.

  The proposed visual analytic model for pass prediction can be used as a basic building block for several open problems like the team formation analysis. The proposed pass prediction model can reveal the ball-passing efficacy of each player and therefore can assist the coach in team formation. Another important research avenue is strategy recommendation. The proposed pass prediction model can be used to exploit the chaotic dynamics of players and possible pass networks can be expressed as a solution of Lyapunov stability equation.

Appendix A

Supplementary for Chapter 2

Let us consider $M^{\theta_b}_{(i+1)}$ be the configuration matrix. The point u in Ω is a row vector of the configuration matrix $M^{\theta_b}_{(i+1)}$.

Lemma 2.1. *According to Eq. 2.4, $S(u) = \left\|(u - M^{\bar{\theta_b}}_{(i+1)})\right\|$, where $M^{\bar{\theta_b}}_{(i+1)}$ is the centroid of all the row vectors of $M^{\theta_b}_{(i+1)}$. S is a shape measure.*

Proof. The shape measure $S : \Omega \to \mathbb{R}^{+} \bigcup \{0\}$ is defined by Dryden and Mardia [1998] as $S(au) = aS(u)$ where a is any positive scalar. To prove the above statement we have to show that for any scalar $a \geq 0$ we will have $S(au) = aS(u)$. As $u \in M^{\theta_b}_{(i+1)}$, so $au \in aM^{\theta_b}_{(i+1)}$. Therefore we can write:

$$\begin{aligned} S(au) &= \left\|au - aM^{\bar{\theta_b}}_{(i+1)}\right\| \\ &= \sqrt{a^2}\left\|u - M^{\bar{\theta_b}}_{(i+1)}\right\| \\ &= |a|S(u). \end{aligned}$$

Thus S is a shape measure. □

Lemma 2.2. $p(X_{0:n}|Y) \propto p(X_{0:1}) \prod_{i=2}^{n} q(X_i; X_{i-2:i-1}) \prod_{u\epsilon(\Omega \bigcap X_n)} l(Y(u))$

Proof.

$$\begin{aligned} p(X_{0:n}|Y) &= \frac{p(X_{0:n} \wedge Y)}{p(Y)} \\ &= \frac{p(Y|X_{0:n})p(X_{0:n})}{p(Y)} \\ &\propto p(Y|X_{0:n})p(X_{0:n}) \end{aligned} \tag{A.1}$$

Now from the assumption of independent events, we have $p(X_{0:n}) = p(X_{0:1}) \prod_{i=2}^{n} q(X_i; X_{i-2:i-1})$ and from Eq. (2.6) we have:

$$p(Y \mid X_{0:n}) = \prod_{u \in \Omega} p_{\text{off}}(Y(u)) \prod_{u \in (\Omega \bigcap X_n)} l(Y(u)). \tag{A.2}$$

Putting A.1 and A.2 two expressions we get:

$$\begin{aligned}
p(X_{0:n}|Y) &= \frac{p(Y|X_{0:n})p(X_{0:n})}{p(Y)} \\
&\propto p(Y|X_{0:n})p(X_{0:n}) \\
&\propto p(X_{0:1})\prod_{i=2}^{n} q(X_i; X_{i-2:i-1}) \prod_{u\in\Omega} p_{\text{off}}(Y(u)) \prod_{u\in(\Omega\bigcap X_n)} l(Y(u)) \\
&\propto p(X_{0:1})\prod_{i=2}^{n} q(X_i; X_{i-2:i-1}) \prod_{u\in(\Omega\bigcap X_n)} l(Y(u))
\end{aligned}$$

So the statement is proved. □

Appendix B

Supplementary for Chapter 3

B.1 Proof of Theorem 3.1

Proof. Let us have the groundtruth F that registers the players of the video frame into the top view model. Given a video frame, the problem is to approximate F. Let, the state-of-the-art method approximates F as F^* defined from Fig. 3.2(b) to 3.2(d). F^* registers players from the video frame into the top-view model with approximation error ϵ i.e. $\|F - F^*\| = \epsilon$, where $\|.\|$ is the Frobenius matrix norm. Given arbitrary F^*, we aim to construct T and H such that (HT) is a better approximation of F. More precisely, given arbitrary F^*, there exists T and H such that (HT) has approximation error less than ϵ. i.e. $\|F - (HT)\| < \epsilon$. The proof is comprised of three steps. First, we construct H and T for a given F which depends on the value of a variable s which will be discussed next. Then we derive the error bound for the proposed approximation (HT) involving s. Finally, the value of s is computed for the given ϵ. Using the value of s, H and T can be computed specifically with approximation error less than ϵ. Next we expound the details.

Construction of H and T: The task of computing a factor form approximation of F can be divided into two steps. The first is to construct a subspace that captures the range of F. The second is to restrict the F to the subspace and compute a standard factorization of the reduced F with the help of H. Next, we discuss how to accomplish the proposed steps.

The first step can be executed with random sampling method [Martinsson and Voronin, 2016]. To understand how randomness works, let us consider $F = B + E$, where B captures the range of F and E is a small perturbation error during groundtruth generation process. Our aim is to obtain a basis of exact rank r that covers as much of the range of B as possible.

Let us consider the dimension of F, B and E are $(a \times b)$. In order to obtain r rank approximation of B, we fix a small number s. Then $(r + s)$ random vectors $\{\alpha_i\}$ can be generated such that:

$$F(\alpha_i) = B(\alpha_i) + E(\alpha_i), \tag{B.1}$$

for $i = 1, ..., (r + s)$. The perturbation E deviates the direction of each $\{\alpha_i\}$ outside the range of B. Therefore, the extra s vectors enhance the chance of spanning the required subspace. Overall the general randomized algorithm to derive the H is comprised of three steps [Halko et al., 2011] as follows.

First, a random $(b \times (r + s))$ matrix $\mathscr{G}$ is generated whose columns are Gaussian vectors.

Thereafter compute $(F\mathscr{G})$. Finally, construct a matrix H whose columns form an orthonormal basis of the range $(F\mathscr{G})$. Once we get the H, then we can compute the other factor (HF). i.e. $F \approx H(HF)$. Considering $T = HF$, we approximate F in factor form of (HT). Next, we compute the bound of approximation error of $\|F - HT\|$. There after, we determine the value of s that is needed to compute H and T so that the approximation error is less than ϵ.

Computing the error bound: We aim to show:

$$E(\|F - H(HF)\|) \leq (1 + \frac{r}{s-1})(\sum_{i=r+1}^{\min(a,b)} \sigma_i^2)$$

where E is the expectation, $\sum_{i=r+1}^{\min(a,b)} \sigma_i^2$ is the theoretically minimal error in approximating F by a matrix of rank r [Halko et al., 2011].

First, consider the singular value decomposition of F as $F = U_1\Sigma_1 V_1^*$, where U_1 is an $(a \times r)$ orthonormal matrix, Σ_1 is a diagonal matrix containing the non negative singular values of F and V_1 is an $(r \times n)$ orthonormal matrix. We call U_1 and V_1 as left unitary factor and right unitary factor respectively. First partition the $\Sigma_1 = [\Sigma_2|\Sigma_3]$, where the Σ_2 and Σ_3 are the diagonal matrix containing the first r and $(b-r)$ singular values respectively. Thereafter, partition $V_1 = [V_2|V_3]$ into blocks containing r and $b-r$ columns respectively. Define $\mathscr{G}_2 = V_2^*\mathscr{G}$ and $\mathscr{G}_3 = V_3^*\mathscr{G}$. Since, V_2 and V_3 are orthonormal, then $\mathscr{G}_2$ and $\mathscr{G}_3$ are also Gaussian. We denote the pseudoinverse of $\mathscr{G}_2$ and $\mathscr{G}_3$ as $\hat{\mathscr{G}}_2$ and $\hat{\mathscr{G}}_3$ respectively. $\mathscr{G}_2$ and $\mathscr{G}_3$ are non overlapping, so they are stochastically independent. Applying Holder's inequality, we can write:

$$E(\|F - H(HF)\|) \leq (E(\|F - H(HF)\|^2))^{1/2} \tag{B.2}$$

It is proved in [Halko et al., 2011] that:

$$E(\|F - H(HF)\|^2) \leq (\|\Sigma_3\|_F^2 + E(\left\|\Sigma_3\mathscr{G}_3\hat{\mathscr{G}}_2)\right\|^2) \tag{B.3}$$

Therefore, using Eq. B.2 and B.3, we can write:

$$E(\|F - H(HF)\|) \leq (\|\Sigma_3\|_F^2 + E(\left\|\Sigma_3\mathscr{G}_3\hat{\mathscr{G}}_2)\right\|^2)^{1/2} \tag{B.4}$$

We are interested in the r ranks of the matrix. Therefore we compute $E(\left\|\Sigma_3\mathscr{G}_3\hat{\mathscr{G}}_2\right\|^2)$ by conditioning on the value of $\mathscr{G}_2$ as follows:

$$E(\left\|\Sigma_3\mathscr{G}_3\hat{\mathscr{G}}_2\right\|^2) = E(E(\left\|\Sigma_3\mathscr{G}_3\hat{\mathscr{G}}_2\right\|^2)|\mathscr{G}_2) \tag{B.5}$$

The Frobenious norm is unitarily invariant. i.e. for any two orthonormal matrices U_1 and V_1, we can write $\|U_1\Sigma_1 V_1\| = \|\Sigma_1\|$. In addition, the distribution of a Gaussian matrix is invariant under orthogonal transformations. Therefore, we can write:

$$\begin{aligned}
E(E(\left\|\Sigma_3\mathscr{G}_3\hat{\mathscr{G}}_2\right\|^2)|\mathscr{G}_2) &= E(E(\Sigma_{jk}(\sigma_{jj}[\mathscr{G}_3]_{jk}[\hat{\mathscr{G}}_2]_{kk})) \\
&= E(\Sigma_{jk}(\sigma_{jj}^2[\hat{\mathscr{G}}_2]_{kk}^2)) \\
&= E(\|\Sigma_3\|_F^2\left\|\hat{\mathscr{G}}_2\right\|^2) \\
&= \|\Sigma_3\|_F^2 E(\left\|\hat{\mathscr{G}}_2\right\|^2) \\
&= \frac{r}{s-1}\|\Sigma_3\|_F^2 \\
&= \frac{r}{s-1}(\sum_{i=r+1}^{\min(a,b)}\sigma_i^2)
\end{aligned} \tag{B.6}$$

Therefore, putting the expression of $E(\left\|\Sigma_3\mathscr{G}_3\hat{\mathscr{G}}_2\right\|^2)$ in the Eq. B.4, we can write:

$$E(\|F - H(HF)\|)) \leq (1 + \frac{r}{s-1})(\sum_{i=r+1}^{\min(a,b)}\sigma_i^2) \tag{B.7}$$

In Eq. B.7, $\sum_{i=r+1}^{\min(a,b)}\sigma_i^2$ is the theoretically minimal error in approximating F by rank r [Halko et al., 2011]. Therefore, the optimal bound is missed by a factor of $(1+\frac{r}{s-1})$. Now, our objective is to determine the value of s for a given ϵ.

Computing the value of s: Our proposed error for factorized approximation is less than ϵ. The target rank r is strictly greater than 1. Therefore from Eq. B.7, we can write:

$$\begin{aligned}
(1+\frac{r}{s-1})(\sum_{i=r+1}^{\min(a,b)}\sigma_i^2) &< \epsilon \\
(1+\frac{r}{s-1}) &< \frac{\epsilon}{(\sum_{i=r+1}^{\min(a,b)}\sigma_i^2)} \\
\frac{r}{s-1} &< \frac{\epsilon}{(\sum_{i=r+1}^{\min(a,b)}\sigma_i^2)} - 1 \\
\frac{r}{s-1} &< \frac{\epsilon - (\sum_{i=r+1}^{\min(a,b)}\sigma_i^2)}{(\sum_{i=r+1}^{\min(a,b)}\sigma_i^2)} \\
\frac{r(\sum_{i=r+1}^{\min(a,b)}\sigma_i^2)}{\epsilon - (\sum_{i=r+1}^{\min(a,b)}\sigma_i^2)} &< s-1 \\
\frac{r(\sum_{i=r+1}^{\min(a,b)}\sigma_i^2)}{\epsilon - (\sum_{i=r+1}^{\min(a,b)}\sigma_i^2)} + 1 &< s
\end{aligned} \tag{B.8}$$

Therefore, we can choose $s = \lceil \frac{r(\sum_{i=r+1}^{\min(a,b)} \sigma_i^2)}{\epsilon - (\sum_{i=r+1}^{\min(a,b)} \sigma_i^2)} + 1 \rceil$, where $\lceil . \rceil$ function gives the least integer greater than or equal to the given input.

The upshot: Putting the value of s we can compute the H and $T = HF$ such that HT is an approximation of F with approximation error less than ϵ. Thus we prove that for a given F and an arbitrary state-of-the-art approximation of F say F^*, we can always find an approximation in factor form of (HT) which is better than F^*. □

B.2 Proof of Lemma 3.1

Proof. From the definition of X^*_κ, we can write

$$\|X^*_\kappa\|_* + \|X^*_\kappa\|_F^2 \leq \|X_\infty\|_* + \|X_\infty\|_F^2. \tag{B.9}$$

From the definition of X_∞ we can write:

$$\|X_\infty\|_* \leq \|X^*_\kappa\|_*. \tag{B.10}$$

From the B.9 and B.10, we can write:

$$\|X^*_\kappa\|_F^2 \leq \|X_\infty\|^2 \tag{B.11}$$

The eq. B.11 implies that X^*_κ is uniformly bounded. Now, the theorem is proved if we can show that any convergent subsequence $\{X^*_{\kappa_i}\}_{i \geq 1}$ must converge to X_∞.

Consider an arbitrary converging subsequence $\{X^*_{\kappa_i}\}$ and set $X_c = \lim_{i \to \infty} X^*_{\kappa_i}$. Since X^*_κ is uniformly bounded, we can write:

$$\lim_{\kappa \to \infty} sup \|X^*_\kappa\| \leq \|X_\infty\|_* \tag{B.12}$$

and

$$\|X^*_\kappa\|_* \leq \lim_{\kappa \to \infty} inf \|X_\infty\|_* \tag{B.13}$$

From Eq. B.12 and B.13, we can write $\lim_{\kappa \to \infty} \|X^*_\kappa\| = \|X_\infty\|_*$, therefore, $\|X_c\|_* = \|X_\infty\|_*$. This shows that X_c is a solution of $\underset{X}{argmin} \|X\|_*$. Now it follows from the definition of X_∞ that $\|X_c\|_F \geq \|X_\infty\|_F$, while we also have $\|X_c\|_F \leq \|X_\infty\|_F$ because of Eq. B.13. Therefore, we conclude that $\|X_c\|_F = \|X_\infty\|_F$ and thus $X_c = X_\infty$ since X_∞ is unique. □

B.3 Proof of Lemma 3.2

Proof. Essentially, we have to prove that:
$O_\kappa(Y) = \underset{X}{argmin} \|X - Y\|_F^2 + \kappa \|X\|_*$. Let us assume $M(X) = \underset{X}{argmin} \|X - Y\|_F^2 + \kappa \|X\|_*$. M is strictly convex. So there exists a unique minimizer of M. Therefore, we need to prove that the minimizer is equal to $O_\kappa(Y)$. To do this, the definition of a subgradient of a convex function is as follows [Cai et al., 2010]: Z is a subgradient of M at X_0, denoted as $Z \in \partial M(X_0)$, if

$$M(X) \geq M(X_0) + \langle Z, (X - X_0) \rangle \tag{B.14}$$

for all X. Now $O_\kappa(Y)$ minimizes M if and only if 0 is a subgradient of the M at the point $O_\kappa(Y)$. i.e.

$$0 \in O_\kappa(Y) - Y + \kappa\partial\|O_\kappa(Y)\|_*, \tag{B.15}$$

where $\partial\|O_\kappa(Y)\|_*$ is the set of subgradients of the nuclear norm. Let X be an arbitrary image and SVD of $X = U\Sigma V^T$. Then we can write [Lewis, 2003] [Watson, 1992]:

$\partial\|X\|_* = \{UV^T + W : W \in \mathbb{R}^{n_1 \times n_2}, U^T W = 0, WV = 0, \|W\|_F \leq 1\}$.

In order to show that $O_\kappa(Y)$ obeys Eq. B.15, decompose the SVD of Y as: $Y = U_3\Sigma_3 V_3^T + U_4\Sigma_4 V_4^T$, where U_3, V_3 (resp. U_4, V_4) are the singular vectors associated with singular values greater than κ (resp. smaller than or equal to κ). With these notations, we have $O_\kappa(Y) = U_3(\Sigma_3 - \kappa I)V_3^T$.

Therefore, $Y - O_\kappa(Y) = \kappa(U_3 V_3^T + W)$, and thus $W = \kappa^{-1}U_4\Sigma_1 V_4^T$.

By definition, $U_0^T W = 0$, $WV_3 = 0$ and since the diagonal elements of $\sum_1$ have magnitudes bounded by κ, we also have $\|W\|_2 \leq 1$. Hence $Y - O_\kappa(Y) \in \kappa\partial\|O_\kappa(Y)\|_*$, which concludes the proof. □

Appendix C

Supplementary for Chapter 4

C.1 Proof of Lemma 4.1

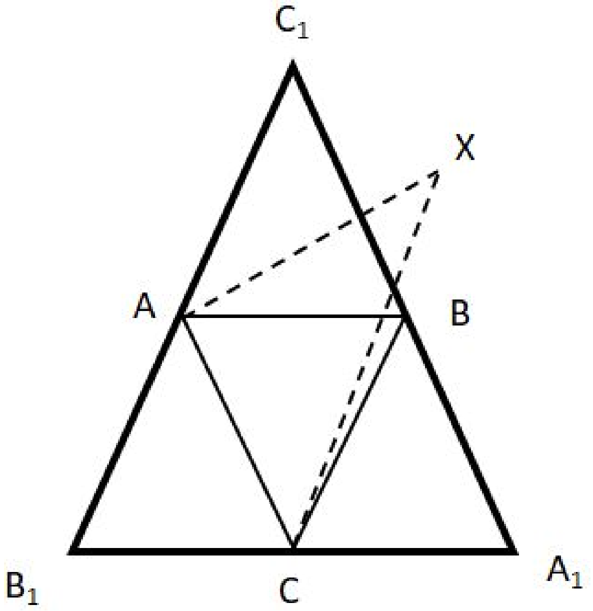


Figure C.1: X is outside the triangle $\Delta A_1B_1C_1$. Which contradicts the assumption of the maximum area triangle ΔABC.

We refer Fig C.1 for the visualization. There are $P \geq 3$ non collinear points given in a plane. So, there are total $\binom{P}{3}$ possible triangles. Among the triangles, consider ΔABC with largest area D. Now, first we will construct $\Delta A_1B_1C_1$. Then we prove that all n points are inside $\Delta A_1B_1C_1$. To construct $\Delta A_1B_1C_1$, we draw lines from the points A, B and C parallel to the lines BC, AC and AB respectively. So, we get $\Delta ABC_1, \Delta AB_1C$ and ΔBCA_1 similar to ΔABC. The ratio of the areas of two similar triangles is equal to the square of the ratio of any pair of their corresponding sides. Therefore, the ratio of the areas of ΔABC_1 and ΔABC is equal to the $\frac{AB^2}{AB^2}$ or 1, as AB is the common corresponding sides of the ΔABC_1 and ΔABC. Consequently, areas of ΔABC_1 and ΔABC are equal.Similarly, we can say that the area of ΔAB_1C, ΔA_1BC and ΔABC are equal. Thus area of all the triangles ΔAB_1C, ΔABC_1, ΔA_1BC and ΔABC are equal to D. Consequently, the area of $\Delta A_1B_1C_1$ becomes $4D$.

Now, we prove that all n points lie inside $\Delta A_1B_1C_1$. Let a point X from given n points is outside $\Delta A_1B_1C_1$. So, we can construct the triangle ΔAXC. Then area of ΔAXC is more than area of ΔABC due to greater height. That contradicts the fact that ΔABC has the largest area. So, all n points are inside $\Delta A_1B_1C_1$.

List of Publications

S. Sanyal. Who will receive the ball? predicting pass recipient in soccer videos. *Journal of Visual Communication and Image Representation*, 78:103190, 2021. 5, 6, 38

S. Sanyal. Tvvs: A top-view visualization system from broadcasting soccer video. *Multimedia Tools and Applications*, 81:33613–33644, 2022. 26

S. Sanyal, A. Kundu, and D. P. Mukherjee. On the (soccer) ball. In *Proceedings of the Tenth Indian Conference on Computer Vision, Graphics and Image Processing*, pages 1–8, 2016. doi: https://doi.org/10.1145/3009977.3010022. 38, 50, 54

References

Dataset. https://www.youtube.com/watch?v=UBJG3FU4DGM, accessed on 23rd Jan 2022a. 18, 24

Dataset1. https://www.youtube.com/watch?v=Wicq4QTlUCk, accessed on 23rd Jan 2022b. 18

M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *Signal Processing, IEEE Transactions on*, 50(2):174–188, 2002. 16, 54

S. Baker and I. Matthews. Lucas-kanade 20 years on: A unifying framework. *International journal of computer vision*, 56(3):221–255, 2004. 28

A. Bialkowski, P. Lucey, P. Carr, Y. Yue, and I. Matthews. Win at home and draw away: Automatic formation analysis highlighting the differences in home and away team behaviors. In *Proceedings of 8th annual MIT sloan sports analytics conference*, pages 1–7. Citeseer, 2014a. 4

A. Bialkowski, P. Lucey, P. Carr, Y. Yue, S. Sridharan, and I. Matthews. Identifying team style in soccer using formations learned from spatiotemporal tracking data. In *2014 IEEE international conference on data mining workshop*, pages 9–14. IEEE, 2014b. 4

A. Bialkowski, P. Lucey, P. Carr, I. Matthews, S. Sridharan, and C. Fookes. Discovering team structures in soccer from spatiotemporal data. *IEEE Transactions on Knowledge and Data Engineering*, 28(10):2596–2605, 2016. 50

J.-F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization*, 20(4):1956–1982, 2010. 38, 70

E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009. 26, 31, 32, 33, 38, 39

P. Carr, Y. Sheikh, and I. Matthews. Point-less calibration: Camera parameters from gradient-based alignment to edge images. In *2012 IEEE Workshop on the Applications of Computer Vision (WACV)*, pages 377–384. IEEE, 2012. 7, 27, 28

S. Chawla, J. Estephan, J. Gudmundsson, and M. Horton. Classification of passes in football matches using spatiotemporal data. *ACM Transactions on Spatial Algorithms and Systems (TSAS)*, 3(2):1–30, 2017a. 7, 51

S. Chawla, J. Estephan, J. Gudmundsson, and M. Horton. Classification of passes in football matches using spatiotemporal data. *ACM Transactions on Spatial Algorithms and Systems (TSAS)*, 3(2):6, 2017b. 7, 50

J. Chen and J. J. Little. Sports camera calibration via synthetic data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 27, 28, 44, 46, 47

P. Cintia, F. Giannotti, L. Pappalardo, D. Pedreschi, and M. Malvaldi. The harsh rule of the goals: Data-driven performance indicators for football teams. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10. IEEE, 2015. 7, 50

F. M. Clemente, M. S. Couceiro, F. M. L. Martins, and R. S. Mendes. Using network metrics to investigate football team players' connections: A pilot study. *Motriz: Revista de Educação Física*, 20(3):262–271, 2014. 7, 50

D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 2, pages 142–149. IEEE, 2000. 18, 20, 21, 22, 23, 24

C. Cotta, A. M. Mora, C. Merelo-Molina, and J. J. Merelo. Fifa world cup 2010: A network analysis of the champion team play. *arXiv preprint arXiv:1108.0261*, 2011. 4

L. Cotta, P. de Melo, F. Benevenuto, and A. Loureiro. Using fifa soccer video game data for soccer analytics. In *Workshop on large scale sports analytics*, 2016. 3

A. Doucet, S. Godsill, and C. Andrieu. On sequential monte carlo sampling methods for bayesian filtering. *Statistics and computing*, 10(3):197–208, 2000. 17

I. L. Dryden and K. V. Mardia. *Statistical shape analysis*, volume 4. J. Wiley Chichester, 1998. 14, 65

E. Dubrofsky and R. J. Woodham. Combining line and point correspondences for homography estimation. In *International Symposium on Visual Computing*, pages 202–213. Springer, 2008. 28

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. 7, 50

C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936. 26

V. M. Farias, W. B. Fernandes, G. G. Bergmann, and E. dos Santos Pinheiro. Relationship between ball possession and match outcome in uefa champions league. *Motricidade*, 16(4): 1–7, 2020. 4

D. Farin, S. Krabbe, W. Effelsberg, et al. Robust camera calibration for sport videos using court models. In *Storage and Retrieval Methods and Applications for Multimedia 2004*, volume 5307, pages 80–91. International Society for Optics and Photonics, 2003. 6, 27

P. Felsen, P. Agrawal, and J. Malik. What will happen next? forecasting player moves in sports videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3342–3351, 2017. 5, 7, 51, 59, 60, 61, 62

J. H. Fewell, D. Armbruster, J. Ingraham, A. Petersen, and J. S. Waters. Basketball teams as strategic networks. *PloS one*, 7(11):e47445, 2012. 7, 50

M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 37

J. Gama, P. Passos, K. Davids, H. Relvas, J. Ribeiro, V. Vaz, and G. Dias. Network analysis and intra-team activity in attacking phases of professional football. *International Journal of Performance Analysis in Sport*, 14(3):692–708, 2014. 7, 50

M. Garon, P.-O. Boulet, J.-P. Doiron, L. Beaulieu, and J.-F. Lalonde. Real-time high resolution 3d data on the hololens. In *2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*, pages 189–191. IEEE, 2016. 64

M. Germann, T. Popa, R. Keiser, R. Ziegler, and M. Gross. Novel-view synthesis of outdoor sport events using an adaptive view-dependent geometry. In *Computer Graphics Forum*, volume 31, pages 325–333. Wiley Online Library, 2012. 25

B. Ghanem, T. Zhang, and N. Ahuja. Robust video registration applied to field-sports video analysis. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2. Citeseer, 2012. 7, 27, 28

S. Giancola, M. Amine, T. Dghaily, and B. Ghanem. Soccernet: A scalable dataset for action spotting in soccer videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018. 56

R. Girshick. Fast r-cnn. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015. doi: 10.1109/ICCV.2015.169. 59

F. R. Goes, M. Kempe, L. A. Meerhoff, and K. A. Lemmink. Not every pass can be an assist: a data-driven model to measure pass effectiveness in professional soccer matches. *Big data*, 7(1):57–70, 2019. 7, 51

B. Gonçalves, D. Coutinho, S. Santos, C. Lago-Penas, S. Jiménez, and J. Sampaio. Exploring team passing networks and player movement dynamics in youth association football. *PloS one*, 12(1):e0171156, 2018. 5, 7, 50

T. U. Grund. Network structure and team performance: The case of english premier league soccer teams. *Social Networks*, 34(4):682–690, 2012. 7, 50

J. Gudmundsson and M. Horton. Spatio-temporal analysis of team sports. *ACM Computing Surveys (CSUR)*, 50(2):22, 2017. 50

A. Gupta, J. J. Little, and R. J. Woodham. Using line and ellipse features for rectification of broadcast hockey video. In *2011 Canadian Conference on Computer and Robot Vision*, pages 32–39. IEEE, 2011. 7, 27, 28

N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53 (2):217–288, 2011. 29, 30, 67, 68, 69

C. G. Harris, M. Stephens, et al. A combined corner and edge detector. In *Alvey vision conference*, volume 15, pages 10–5244. Citeseer, 1988. 37

R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 37

J.-B. Hayet and J. Piater. On-line rectification of sport sequences with moving cameras. In *Mexican International Conference on Artificial Intelligence*, pages 736–746. Springer, 2007. 7, 28

J.-B. Hayet, J. Piater, and J. Verly. Robust incremental rectification of sports video sequences. In *British Machine Vision Conference (BMVC'04)*, pages 687–696. Citeseer, 2004. 7, 28

R. Hess and A. Fern. Improved video registration using non-distinctive local image features. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 28

J. Hobbs, P. Power, L. Sha, and P. Lucey. Quantifying the value of transitions in soccer via spatiotemporal trajectory clustering. In *MIT Sloan Sports Analytics Conference*, 2018. 6, 27

N. Homayounfar, S. Fidler, and R. Urtasun. Sports field localization via deep structured models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5212–5220, 2017. 3, 4, 7, 27, 28, 29, 37, 38, 41, 43, 44

Y. Huang, J. Llach, and C. Zhang. A method of small object detection and tracking based on particle filters. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE, 2008. 6, 10

H. Janetzko and S. I. Fabrikant. Conveying Uncertainty in Visual Cluster Representations of Soccer Player Trajectories. In *The 28th International Cartographic Conference (ICC 2017), Abstract*, 2017. 3

X. Jia, H. Lu, and M.-H. Yang. Visual tracking via adaptive structural local sparse appearance model. In *Computer vision and pattern recognition (CVPR), 2012 IEEE Conference on*, pages 1822–1829. IEEE, 2012. 20

W. Jiang, J. C. G. Higuera, B. Angles, W. Sun, M. Javan, and K. M. Yi. Optimizing through learned errors for accurate sports field registration. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 201–210, 2020. 27, 28, 29, 44, 46, 47

T. Kawasaki, K. Sakaue, R. Matsubara, and S. Ishizaki. Football pass network based on the measurement of player position by using network theory and clustering. *International Journal of Performance Analysis in Sport*, 19(3):381–392, 2019. 5

A. Kendall, M. Grimes, and R. Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international conference on computer vision*, pages 2938–2946, 2015. 28

H. Kim and K. S. Hong. Soccer video mosaicing using self-calibration and line tracking. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, volume 1, pages 592–595. IEEE, 2000. 6, 27

A. S. Lewis. The mathematics of eigenvalue optimization. *Mathematical Programming*, 97 (1-2):155–176, 2003. 71

D. Link and M. Hoernig. Individual ball possession in soccer. *PloS one*, 12(7):e0179953, 2017. 7, 50, 57

J. López Peña and H. Touchette. A network theory analysis of football strategies. *arXiv*, pages arXiv–1206, 2012. 7, 50

D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 36

P. Lucey, A. Bialkowski, M. Monfort, P. Carr, and I. Matthews. quality vs quantity: Improved shot prediction in soccer using strategic features from spatiotemporal data. In *Proc. 8th annual mit sloan sports analytics conference*, pages 1–9, 2014. 1

D. Lusher, G. Robins, and P. Kremer. The application of social network analysis to team sports. *Measurement in physical education and exercise science*, 14(4):211–224, 2010. 7, 50

S. Mahendran, H. Ali, and R. Vidal. Convolutional networks for object category and 3d pose estimation from 2d images. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 28

A. Maksai, X. Wang, and P. Fua. What players do with the ball: A physically constrained interaction modeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 972–981, 2016. 50

P.-G. Martinsson and S. Voronin. A randomized blocked algorithm for efficiently computing rank-revealing factorizations of matrices. *SIAM Journal on Scientific Computing*, 38(5): S485–S507, 2016. 29, 67

T. B. Moeslund, G. Thomas, and A. Hilton. *Computer Vision in Sports*. Springer, 2015. 10

K. Okuma, J. J. Little, and D. G. Lowe. Automatic rectification of long image sequences. In *Asian Conference on Computer Vision*, volume 9, 2004. 7, 28

V. Pallavi, J. Mukherjee, A. K. Majumdar, and S. Sural. Ball detection from broadcast soccer videos using static and dynamic features. *Journal of Visual Communication and Image Representation*, 19(7):426–436, 2008a. 10

V. Pallavi, J. Mukherjee, A. K. Majumdar, and S. Sural. Graph-based multiplayer detection and tracking in broadcast soccer videos. *IEEE Transactions on Multimedia*, 10(5):794–805, 2008b. 10

L. Pappalardo, P. Cintia, A. Rossi, E. Massucco, P. Ferragina, D. Pedreschi, and F. Giannotti. A public data set of spatio-temporal match events in soccer competitions. *Scientific data*, 6(1):1–15, 2019. 56

P. Pérez, A. Blake, and M. Gangnet. Jetstream: Probabilistic contour extraction with particles. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 2, pages 524–531. IEEE, 2001. 10, 13

C. Perin, R. Vuillemot, and J.-D. Fekete. Soccerstories: A kick-off for visual soccer analysis. *IEEE transactions on visualization and computer graphics*, 19(12):2506–2515, 2013. 2

S. A. Pettersen, D. Johansen, H. Johansen, V. Berg-Johansen, V. R. Gaddam, A. Mortensen, R. Langseth, C. Griwodz, H. K. Stensland, and P. Halvorsen. Soccer video and player position dataset. In *Proceedings of the 5th ACM Multimedia Systems Conference*, pages 18–23, 2014. 55

P. Power, H. Ruiz, X. Wei, and P. Lucey. Not all passes are created equal: Objectively measuring the risk and reward of passes in soccer from tracking data. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1605–1613, 2017. 7, 51

Prozone. Prozone. `http://prozonesports.stats.com/`, link accessed March 2022. Accessed: 2022-3-6. 27

V. Ramanathan, J. Huang, S. Abu-El-Haija, A. Gorban, K. Murphy, and L. Fei-Fei. Detecting events and key actors in multi-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3043–3053, 2016. 50

M. G. I. Rathod and M. D. A. Nikam. Review on event retrieval in soccer video. *Int. J. Comput. Sci. Inf. Technol*, 5:5601–5605, 2014. 1

J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *arXiv*, 2018. 35, 58, 60

C. Reep and B. Benjamin. Skill and chance in association football. *Journal of the Royal Statistical Society. Series A (General)*, 131(4):581–585, 1968. 2

R. Rein and D. Memmert. Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. *SpringerPlus*, 5(1):1–13, 2016. 4, 50

R. Rein, D. Raabe, and D. Memmert. "which pass is better?" novel approaches to assess passing effectiveness in elite soccer. *Human movement science*, 55:172–181, 2017. 7, 51

D. Sacha, M. Stein, T. Schreck, D. A. Keim, O. Deussen, et al. Feature-driven visual analytics of soccer data. In *2014 IEEE conference on visual analytics science and technology (VAST)*, pages 13–22. IEEE, 2014. 2

D. Sacha, F. Al-Masoudi, M. Stein, T. Schreck, D. A. Keim, G. Andrienko, and H. Janetzko. Dynamic visual abstraction of soccer movement. In *Computer Graphics Forum*, volume 36, pages 305–315. Wiley Online Library, 2017. 3

S. Sanyal. Who will receive the ball? predicting pass recipient in soccer videos. *Journal of Visual Communication and Image Representation*, page 103190, 2021. 5, 6, 38

S. Sanyal. Tvvs: A top-view visualization system from broadcasting soccer video. *Multimedia Tools and Applications*, pages 1–32, 2022. 26

S. Sanyal, A. Kundu, and D. P. Mukherjee. On the (soccer) ball. In *Proceedings of the Tenth Indian Conference on Computer Vision, Graphics and Image Processing*, pages 1–8, 2016. 38, 50, 54

S. Sarkar, A. Chakrabarti, and D. Prasad Mukherjee. Generation of ball possession statistics in soccer using minimum-cost flow network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 4, 50

L. Sha, P. Lucey, Y. Yue, X. Wei, J. Hobbs, C. Rohlf, and S. Sridharan. Interactive sports analytics: An intelligent interface for utilizing trajectories for interactive sports play retrieval and analytics. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 25 (2):1–32, 2018. 6, 27

L. Sha, J. Hobbs, P. Felsen, X. Wei, P. Lucey, and S. Ganguly. End-to-end camera calibration for broadcast videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13627–13636, 2020. 4, 27, 44, 46, 47

R. A. Sharma, B. Bhat, V. Gandhi, and C. V. Jawahar. Automated top view registration of broadcast football videos. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 305–313, 2018. 25, 27, 42, 44, 47

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 59

W. Spearman, A. Basye, G. Dick, R. Hotovy, and P. Pop. Physics—based modeling of pass probabilities in soccer. In *Proceeding of the 11th MIT Sloan Sports Analytics Conference*, 2018. 5, 7, 51, 56, 59, 60, 61, 62

Stathlete. Manual. https://stathlete.ie/, link accessed March 2022. Accessed: 2022-2-3. 27

M. Stein, H. Janetzko, T. Breitkreutz, D. Seebacher, T. Schreck, M. Grossniklaus, I. D. Couzin, and D. A. Keim. Director's cut: Analysis and annotation of soccer matches. *IEEE computer graphics and applications*, 36(5):50–60, 2013. 1

M. Stein, J. Häußler, D. Jäckle, H. Janetzko, T. Schreck, and D. A. Keim. Visual soccer analytics: Understanding the characteristics of collective team movement based on feature-driven analysis and abstraction. *ISPRS International Journal of Geo-Information*, 4(4): 2159–2184, 2015. 2

M. Stein, H. Janetzko, T. Breitkreutz, D. Seebacher, T. Schreck, M. Grossniklaus, I. D. Couzin, and D. A. Keim. Director's cut: Analysis and annotation of soccer matches. *IEEE computer graphics and applications*, 36(5):50–60, 2016. 51

M. Stein, H. Janetzko, T. Schreck, and D. A. Keim. Tackling similarity search for soccer match analysis: multimodal distance measure and interactive query definition. *IEEE computer graphics and applications*, 39(5):60–71, 2019. 2

T. Taki, J.-i. Hasegawa, and T. Fukumura. Development of motion analysis system for quantitative evaluation of teamwork in soccer games. In *Proceedings of 3rd IEEE international conference on image processing*, volume 3, pages 815–818. IEEE, 1996. 7, 51

B. Tekin, S. N. Sinha, and P. Fua. Real-time seamless single shot 6d object pose prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 292–301, 2018. 28

P. H. Torr and A. Zisserman. Mlesac: A new robust estimator with application to estimating image geometry. *Computer vision and image understanding*, 78(1):138–156, 2000. 37

T. Von Landesberger, S. Bremm, T. Schreck, and D. W. Fellner. Feature-based automatic identification of interesting data segments in group movement data. *Information Visualization*, 13(3):190–212, 2014. 2

F. Wang, L. Sun, B. Yang, and S. Yang. Fast arc detection algorithm for play field registration in soccer video mining. In *2006 IEEE International Conference on Systems, Man and Cybernetics*, volume 6, pages 4932–4936. IEEE, 2006. 6, 27

T. Watanabe, M. Haseyama, and H. Kitajima. A soccer field tracking method with wire frame model from tv images. In *2004 International Conference on Image Processing, 2004. ICIP'04.*, volume 3, pages 1633–1636. IEEE, 2004. 6, 27

G. A. Watson. Characterization of the subdifferential of some matrix norms. *Linear algebra and its applications*, 170:33–45, 1992. 71

Weblink. Freed. https://www.intel.com/content/www/us/en/sports/technology/true-view.html, accessed March 2022. Accessed: 2022-1-3. 25, 27

X. Wei, L. Sha, P. Lucey, S. Morgan, and S. Sridharan. Large-scale analysis of formations in soccer. In *2013 international conference on digital image computing: techniques and applications (DICTA)*, pages 1–8. IEEE, 2013. 4

P.-C. Wen, W.-C. Cheng, Y.-S. Wang, H.-K. Chu, N. C. Tang, and H.-Y. M. Liao. Court reconstruction for camera calibration in broadcast basketball videos. *IEEE transactions on visualization and computer graphics*, 22(5):1517–1526, 2015. 28

Y. Wu, J. Lim, and M.-H. Yang. Online object tracking: A benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2411–2418, 2013. 19

Y. Wu, X. Xie, J. Wang, D. Deng, H. Liang, H. Zhang, S. Cheng, and W. Chen. Forvizor: Visualizing spatio-temporal team formations in soccer. *IEEE transactions on visualization and computer graphics*, 25(1):65–75, 2018. 4, 50

Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017. 28

J. Xing, H. Ai, L. Liu, and S. Lao. Multiple player tracking in sports video: A dual-mode two-way bayesian inference approach with progressive observation modeling. *Image Processing, IEEE Transactions on*, 20(6):1652–1667, 2011. 6, 10

C. Xu, Y.-F. Zhang, G. Zhu, Y. Rui, H. Lu, and Q. Huang. Using webcast text for semantic event detection in broadcast sports video. *IEEE Transactions on Multimedia*, 10(7):1342–1355, 2008. 2

A. Yamada, Y. Shirai, and J. Miura. Tracking players and a ball in video image sequence and estimating camera parameters for 3d interpretation of soccer games. In *Object recognition supported by user interaction for service robots*, volume 1, pages 303–306. IEEE, 2002. 7, 27

X. Yu, C.-H. Sim, J. R. Wang, and L. F. Cheong. A trajectory-based ball detection and tracking algorithm in broadcast tennis video. In *Image Processing, 2004. ICIP'04. 2004 International Conference on*, volume 2, pages 1049–1052. IEEE, 2004. 6, 10

Z. Yue, H. Broich, and J. Mester. Statistical analysis for the soccer matches of the first bundesliga. *International Journal of Sports Science & Coaching*, 9(3):553–560, 2014. 2

E. Zhan, S. Zheng, Y. Yue, L. Sha, and P. Lucey. Generating multi-agent trajectories using programmatic weak supervision. *arXiv preprint arXiv:1803.07612*, 2018. 6, 27