MODEL FIELD PARTICLES WITH POSITIONAL APPEARANCE LEARNING FOR SPORTS PLAYER TRACKING

A DISSERTATION SUBMITTED TO

THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

COMPUTER ENGINEERING

By

Sermetcan Baysal

June 2016

MODEL FIELD PARTICLES WITH POSITIONAL APPEARANCE LEARNING FOR SPORTS PLAYER TRACKING

By Sermetcan Baysal

June 2016

We certify that we have read this dissertation and that in our opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

Selim Aksoy(Advisor)

Pınar Duygulu Şahin(Co-Advisor)

Aydın Alatan

Uğur Güdükbay

Çiğdem Gündüz Demir

Selen Pehlivan

Approved for the Graduate School of Engineering and Science:

Levent Onural
Director of the Graduate School

ABSTRACT

MODEL FIELD PARTICLES WITH POSITIONAL APPEARANCE LEARNING FOR SPORTS PLAYER TRACKING

Sermetcan Baysal
Ph.D. in Computer Engineering
Advisor: Selim Aksoy
Co-Advisor: Pınar Duygulu Şahin
June 2016

Tracking multiple players is crucial to analyzing sports videos in real time. Yet, illumination variations, background clutter, frequent occlusions among players who look similar in low-resolution, and non-linear motion patterns of the targets make sports player tracking difficult. Particle-filtering based approaches have been utilized for their ability in tracking under occlusion and rapid motions. Unlike the common practice of choosing particles on targets, we introduce the notion of shared particles densely sampled at fixed positions on the model field. Likelihoods of being on different particles are calculated for the targets using the proposed combined appearance and motion model. After globally distributing particles among the tracks, particles are weighted using an appearance model with a player detection score, and the track locations are updated by the weighted combination of the particles. This enables encapsulating the interactions among the targets in the state-space model and tracking players through challenging occlusions. We further introduce collective motion model and positional appearance learning to recover lost players and detect identity switches among the tracks. The proposed algorithm is embedded into a real player tracking system. Complete steps of the system are described and the proposed approach is evaluated on large-scale video. Experimental results show that the proposed tracker performs better than standard particle filtering and the state-of-the-art single-object trackers by losing less number of tracks and preserving more identities. Moreover, the proposed approach achieves a higher tracking accuracy with lower error rates on a publicly available soccer tracking dataset when compared to the previous methods.

Keywords: Sports video analysis, Sports player tracking, Multiple object tracking, Model field particles, Positional appearance learning, Collective motion model.

ÖZET

SPORCU TAKİBİ İÇİN SAHA MODELİ PARÇACIKLARI VE POZİSYON TABANLI GÖRÜNÜM ÖĞRENİMİ

Sermetcan Baysal
Bilgisayar Mühendisliği, Doktora
Tez Danışmanı: Selim Aksoy
Eş Tez Danışmanı: Pınar Duygulu Şahin
Haziran 2016

Çoklu oyuncu takibi, gerçek zamanlı spor video analizi için çok önemlidir. Ancak, ortam ışığındaki değişkenlik, arka plan karışıklığı, benzer görünümlü oyuncuların düşük çözünürlükte sıkça birbirlerini engellemeleri, hedeflerin hızlı ve doğrusal olmayan hareketleri sporda oyuncu takibini zorlaştırmaktadır. Hedefleri görünüm kapanması ve hızlı hareket altında da takip edebilme yeteneklerinden dolayı parçacık filtresini temel alan yöntemlerden sıkça faydalanılmaktadır. Bu çalışmada, parçacıkları hedefler üzerinden seçen yaygın kullanımdan farklı olarak, parçacıkları bir saha modeli üzerindeki sabit noktalardan yoğun olarak örnekleme kavramı sunulmaktadır. Hedeflerin saha parçacıkları üzerinde olma olasılıkları, birleşik görünüm ve hareket modeli ile hesaplanmaktadır. Parçacıklar hedefler arasında dağıtıldıktan sonra, tüm parçacıklara oyuncu algılama skoru kullanan bir görünüm modeli ile ağırlıklar atanmakta ve bu ağırlıklar kullanılarak hedeflerin yeri güncellenmektedir. Böylece, oyuncular arasındaki etkileşim yöntem içinde kapsanmakta ve oyuncular zorlu koşullar altında takip edilebilmektedir. Ayrıca, sunulan toplu hareket modeli ve pozisyon tabanlı görünüm öğrenimi ile kaybolan oyuncular geri kazanılmakta ve hedefler arasındaki kimlik değişimleri algılanmaktadır. Sunulan yöntem gerçek bir futbolcu takip sisteminin içine gömülmüştür. Bu sistemin tüm adımları anlatılmakta ve yöntem büyük ölçekli görüntü verisi üzerinde değerlendirilmektedir. Deneysel sonuçlar, sunulan yöntemin, standart parçacık filtresi ve tek nesne takibi yöntemlerine göre daha az hedef kaybettiğini ve daha fazla hedef kimliği koruduğunu göstermiştir. Dahası, sunulan yöntem herkese açık bir veri kümesi üzerinde, önceki çalışmalardan daha başarılı sonuçlar almıştır.

Anahtar sözcükler: Spor video analizi, Sporcu takibi, Çoklu nesne takibi, Saha modeli parçacıkları, Pozisyon tabanlı görünüm öğrenimi, Toplu hareket modeli.

Acknowledgement

At his Stanford University commencement speech, Steve Jobs said “Your work is going to fill a large part of your life, and the only way to be truly satisfied is to do what you believe is great work. And the only way to do great work is to love what you do.” I consider myself among the lucky ones who found what they love and do great work. I have enjoyed every moment of my last five years researching on sports video analysis and applying my research in my sports technology company to develop software that is used around the world.

First and foremost, I owe my deepest gratitude to my advisor Pınar Duygulu Şahin, for letting me do what I love, for her encouragement, guidance and support throughout my studies. I am also thankful to my new advisor Selim Aksoy for accepting to work with me and helping me in completing my thesis. I am grateful to the members of my thesis committee, Uğur Güdükbay, Çiğdem Gündüz Demir, Selen Pehlivan for accepting to read and review my thesis and for their insightful comments. I would like to make a special reference to Aydın Alatan, who has been involved in my work since my master’s thesis, for always giving a hand in those times when I struggled in my research.

Sentio Sports Technology was the intermediary between my academic research and the sports industry. It was very challenging to both conduct research and apply it in a real-world application. But it was rewarding at the end to witness that my work is being recognized and appreciated by the sports community. I would like to thank Serdar Alemdar for walking with me in this journey, always being supportive and positive, and making me believe in myself. I offer my regards and blessings to all my colleagues and friends from Sentio, who have supported me in any aspect or contributed to my research. It was great working with Hande Alemdar, Canberk Bacı, Emre Erçin, Hakan Özgür, Serhat Kurtuluş, Mustafa Alparslan, Fırat Hocaoğlu, Ervin Domazet, Münir Sali, Kadir Korkmaz, and with those whom I forgot to mention.

It was great to have met all my classmates, my officemates, instructors and faculty members in Bilkent. Especially, I would like to mention Eren Gölge, Cağrı Toraman, Hüseyin Gökhan Akçay, Can Fahrettin Koyuncu, İstemi Bahçeci, and Shatlyk Ashyralyyev for all the good memories and also being there and supporting me in my thesis defense.

Last but not the least, I would like to very much thank to İpek Lale and to my parents İnci and Ayhan Baysal for always being there for me, trusting in me, and making me feel comfortable at all times. None of this would have been possible without their love.

My grandmother always wanted to see me finish my work and graduate, but sadly she couldn't make it. I dedicate my thesis to my grandmother, to my mom and to my dad...

Contents

List of Figures

List of Tables

List of Symbols

Symbols related to the concept of model field particles

$\mathbf{S}$	set of densely sampled model field particles with size M
$\mathbf{s}^m$	m-th particle in $\mathbf{S}$, where $1 \leq m \leq \mathrm{M}$
$\mathbf{q}^m$	unique position (x, y) of $\mathbf{s}^m$ on the ground plane
B^m	corresponding bounding box of $\mathbf{s}^m$ on the image plane
$\mathbf{a}^m$	appearance model of $\mathbf{s}^m$
e^m	likelihood of $\mathbf{s}^m$ containing a player
$\mathbf{S}'$	subset of particles that may contain a player $\mathbf{S}' \subset \mathbf{S}$
$\mathbf{S}^+$	subset of particles containing a player $\mathbf{S}^+ \subset \mathbf{S}$
$I[B^m]$	image patch described by the B^m
$hog(I[B^m])$	gradient features of the given image patch
$\mathrm{hSVM}(hog)$	classify given gradient vector using global player detector

Symbols related to model field construction

$\mathbf{q}_d$	a distorted point (x_d, y_d) on the image plane
$\mathbf{q}_{image}$	an undistorted point (x_{image}, y_{image}) on the image plane
H	homography matrix to transform $\mathbf{q}_m$ to $\mathbf{q}_{image}$
κ	radial distortion coefficient
(c_x, c_y)	center coordinates of an image
L_j	field boundary lines on the image, where $1 \leq j \leq 4$
$\mathbf{u}_j^i$	i-th calibration point marked on line L_j
$\mathbf{u}_{bottom}$	bottom point of goal post on the image plane
$\mathbf{u}_{top}$	top point of goal post on the image plane
T_{goal}	fixed reference height of the goal post in real-world (in meters)
T_{player}	fixed height of B^m in real-world (in meters)
$L_{horizon}$	imaginary horizon line on the image plane
h_{cam}	height of the camera above the ground in real-world (in meters)
h_{player}	height of B^m on the image plane (in pixels)

Symbols related to multi-player tracking

$\mathbf{X}_t$	set of track states at time t consisting of N tracks
$\mathbf{x}_t^n$	n-th track state in $\mathbf{X}_t$ at time t, where $1 \leq n \leq$ N
$\mathbf{p}_t^n$	predicted position (x, y) of $\mathbf{x}_t^n$ on ground plane at time t
$\mathbf{v}_t^n$	velocity of $\mathbf{x}_t^n$ at time t
$\mathbf{b}^n$	appearance model of $\mathbf{x}_t^n$
$p(\mathbf{x}_t^n \mid \mathbf{x}_{t-1}^n)$	next state prediction at time t, given previous state $\mathbf{x}_{t-1}^n$
F	state transition model
ω_t	process noise representing acceleration at time t
r_{max}	maximum distance (in meters) a track can travel in Δt
$f(\mathbf{x}^n)$	subset of particles in $\mathbf{S}^+$associated with track $\mathbf{x}^n$
$p(\mathbf{s}^m \mid \mathbf{x}^n)$	likelihood of track $\mathbf{x}_t^n$ being on particle $\mathbf{s}^m$
$w(\mathbf{x}^n, \mathbf{s}^m)$	weight of particle $\mathbf{s}^m$ among those associated with $\mathbf{x}^n$
$\mathbf{p}_{observed}^n$	observed position (x, y) of $\mathbf{x}_t^n$ on ground plane at time t
ϵ	measurement noise of observation
$g(\mathbf{s}^m)$	subset of tracks in $\mathbf{X}_t$ claiming to be on the particle $\mathbf{s}^m$
$p_{\text{<model>}}(\mathbf{s}^m \mid \mathbf{x}^n)$	likelihood of $\mathbf{x}^n$ being on $\mathbf{s}^m$ with respect to <model>
$d_{color}(\mathbf{b}^n, \mathbf{a}^m)$	similarity of color histograms $\mathbf{b}^n$ and $\mathbf{a}^m$
$d_{motion}(\mathbf{p}^n, \mathbf{q}^m)$	distance between points $\mathbf{p}^n$ and $\mathbf{q}^m$ on the ground plane
$\delta(d)$	normal distribution function with zero mean
σ_{motion}	standard deviation of δ function

Symbols related to player identification

$\mathbf{X}_k$	subset of tracks assigned to team k, where $k = 1, 2$
$\mathbf{x}_k^n$	n-th track state in $\mathbf{X}_k$, where $1 \leq n \leq$ N and N $= \lvert\mathbf{X}_k\rvert$
Y_k	set of player identities in team k, where $k = 1, 2$
y_k^i	i-th player identity in Y_k, where $1 \leq i \leq 10$
$\mathbf{pp}_k^i$	estimated position (x, y) of player identity y_k^i
$cost(\mathbf{x}_k^n \leftarrow y_k^i)$	cost of assigning player identity tag y_k^i to track $\mathbf{x}_k^n$
$tag(\mathbf{x}_k^n) \leftarrow y_k^i$	assign player identity tag to y_k^i to track $\mathbf{x}_k^n$
r_{search}	search radius (in meters) for a lost player
$d(\mathbf{x}_k^n, \mathbf{x}_k^i)$	distance between tracks $\mathbf{x}_k^n$ and $\mathbf{x}_k^i$ on ground plane
$\theta(\mathbf{x}_k^n, \mathbf{x}_k^i)$	angle between tracks $\mathbf{x}_k^n$ and $\mathbf{x}_k^i$ on ground plane
$rpd(\mathbf{x}_k^n)$	relative position descriptor of track $\mathbf{x}_k^n$
r_{pos}	distance threshold for calculating relative position descriptor
$ad(\mathbf{x}_k^n)$	appearance descriptor of track $\mathbf{x}_k^n$
$\text{aSVM}_k(ad)$	classify given appearance descriptor, output label $\in Y_k$
$\text{pSVM}_k(rpd)$	classify given relative position descriptor, output label $\in Y_k$
Φ	global set of recent classifications
$p_{tag}(y_k^i \mid \mathbf{x}_k^n)$	probability of $\mathbf{x}_k^n$ being y_k^i, derived from Φ

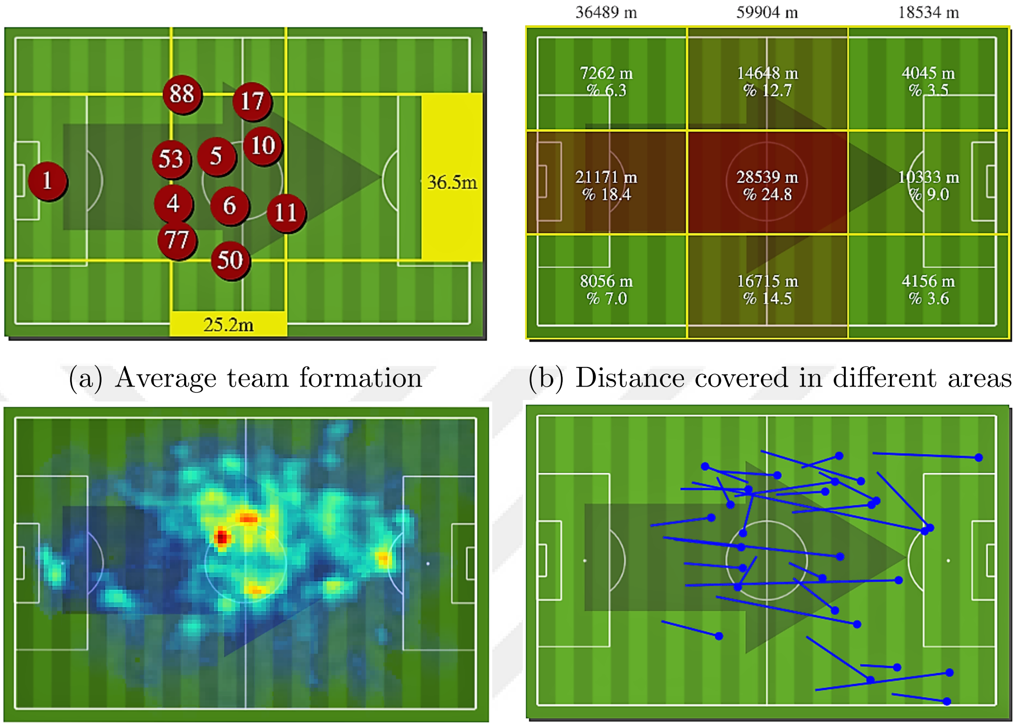
Chapter 1

Introduction

1.1 Motivation and Challenges

Recent advancements in technology has made a great impact on sports. A wide spectrum of applications has been introduced to offer: analysis of sports performance to improve the quality of feedback given to player/athletes, support for referees in making better decisions, automatic extraction of highlights or moments of interest from game videos, intelligent broadcast cameras that can operate automatically (see [1] for a detailed list of applications).

Data collection constructs the basis of all sport technologies. Currently, videos are the most popular way of collecting sports data since they encapsulate rich information, are available to everyone as television broadcast footage, and can be obtained by placing a few cameras in the stadium or even by personal mobile phone cameras of the audience. This has led many computer vision researchers to work on sports video analysis, especially on soccer (referred to as football in most of the world), since it is the most popular sport worldwide having near 260 million players, 300,000 clubs with fan participation in the billions (FIFA Big Count Survey in 2006 [2]).



(a) Average team formation

(b) Distance covered in different areas

(c) Heatmap of a player

(d) Sprints of a player

Figure 1.1: Samples of analysis data provided in real-time by Sentio Sports Analytics [3] to the soccer teams. Tracking data are extracted using the proposed multi-player tracking system.

Team/player performance measurement systems have a solid value proposition because of their potential to reveal aspects of the game that are not obvious to the human eye. Such systems can measure the distance covered by players, speed of movement, number of sprints, and players' relative positioning with respect to others (see Figure 1.1 for example illustrations). These data are then used in individual player performance evaluation, fatigue detection, assessment of team's tactical performance and analysis of the opponents.

Accurate tracking of multiple soccer players in real time, is the key aspect of extracting metrics for performance evaluation, and requires detecting players on video, finding their positions at regular intervals, and linking spatio-temporal data to extract trajectories. However, multiple player tracking is a non-trivial task due to various challenges. Unlike vehicles or pedestrians, which have relatively predictable motion patterns, soccer players try to confuse each other with unexpected changes in velocity. Moreover, players look almost identical because of their jerseys and they are frequently involved in possession challenges and tackles, where they can be occluded by a peer, resulting in tracking ambiguities. Last but not least, environmental conditions can also negatively affect the process of player segmentation. Light changes rapidly during cloudy weather, dark and long player shadows fall on the field in sunny weather, and electronic billboards continuously blink around the stadium during night matches. All of these factors can make it difficult to locate and track players on the field.

1.2 Overview and Contributions

As described in [4], it is common to encapsulate the descriptive information of a soccer match (such as player position, velocity and appearance) into states at each time frame to model the game as a collection of temporal states. Then, the multiple player tracking problem can be perceived as a stochastic process, where the objective is to estimate the state of the game based on the previous observations. Some previous methods use a joint representation of the target space and a unified observation model for all players resulting in a huge state-space. A wrong estimation of a single player may negatively affect the whole state and make the formulation intractable. In contrast, other methods decouple the player states and employ a separate tracker for each target. Although these approaches are efficient and simpler to formulate, they can neither grasp the global state of the game nor the relations among the players, resulting in the well-known problem of identity hijacking.

As a solution, we propose a robust method to accurately track multiple soccer players that combines the relative efficiency of employing separate probabilistic trackers with the effectiveness of joint-state models. Unlike the common practice of choosing particles on the targets, we introduce the concept of model field particles. The ground plane is spanned by densely sampled particles representing the possible positions that the players can occupy. Players are tracked separately on the model field and the position of a player is estimated by a set of neighbor particles. The overall state of the game and the interactions among the players are handled by distributing and sharing particles among the tracks. Distribution is made by globally evaluating the likelihood of the tracks being on the particles. The concept of model field particles, implicitly resolve occlusions and track targets with almost identical appearances, since an occlusion may only occur on the image plane and tracks cannot occlude each other or be on top of each other on the ground plane.

The other contributions, complementary to the concept of model field particles, presented in this thesis are as follows:

i. We propose a hybrid track-to-particle likelihood formulation in which a combined color and motion model is used for distributing particles among tracks, and a combined color and global soccer player appearance model is used for estimating final track positions.

ii. We present an approach for locating players on the model field robust to challenging illumination and environmental conditions.

iii. We describe a method to estimate the position of lost players using a regional collective motion model and an optimal assignment-based algorithm to recover from track losses in the short-term.

iv. Last but not least, we present a positional appearance learning model to detect incorrect identities on the tracks and initialize new observations with correct player identities in the long-term.

The proposed approach has been implemented and embedded in a real-time, two-camera, soccer-player tracking system, called the *Sentioscope*, and has been continuously tested and evolved in near 440 professional soccer league matches tracking players in 12 different countries covering a total distance of 100,000 km.

Experimental results demonstrate that our methodology is better at preserving identities of the players during occlusions, and is more suitable for multiple object tracking with similar appearances such as in team sports when compared to the standard particle filtering methods and the state-of-the-art single-object trackers. Moreover, our approach shows a favorable performance on a publicly available tracking dataset when compared to recent multi-player tracking methods. The overview of our approach is depicted in Figure 1.2.

1.3 Organization of the Thesis

The remainder of this thesis is organized as follows:

Chapter 2 presents a review of recent studies related with sports player tracking, and provides a discussion on comparison of our approach with the related studies.

Chapter 3 describes the proposed methodology. It explains the concept of model field particles; provides details on constructing model field particles, detecting players, and tracking multiple players. Chapter is concluded by sections on short-term player identity recovery, and positional appearance learning to detect and correct player identity mismatches.

Chapter 4 evaluates the performance of different aspects of our approach on a dataset collected from a professional soccer match. It further compares our tracker with state-of-the-art single-object trackers, and uses a publicly available dataset to compare our methodology with the recent studies in sports player tracking.

Chapter 5 concludes the thesis by giving a summary and discussion of our approach and describes possible future extensions.

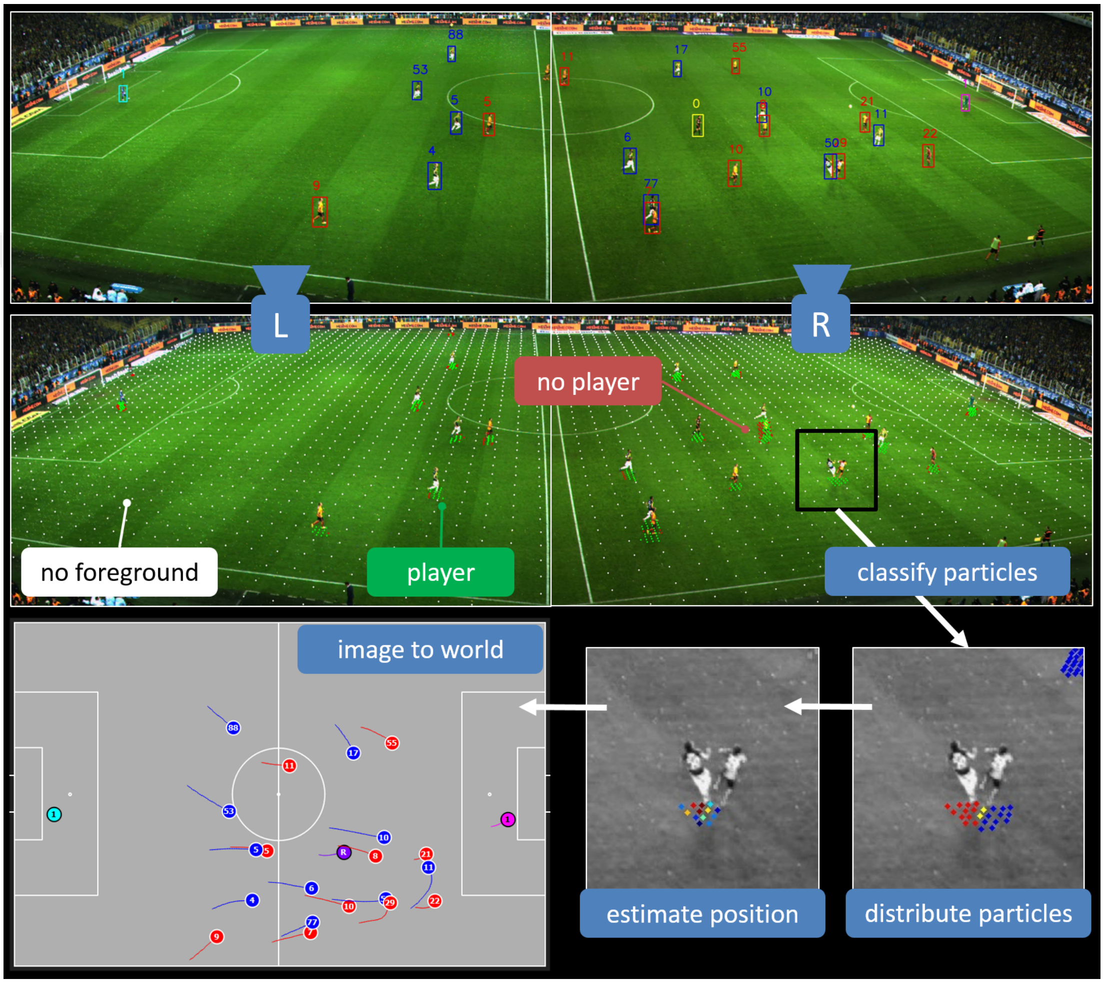


Figure 1.2: Overview of our approach (best viewed in color). *Top row*: Two cameras configured to view the left and right half of the soccer field respectively. *Middle row*: Sparse illustration of model field particles. Particles having no foreground pixels are shown in white, candidate particles having foreground regions are shown in red, and candidate particles that are positively classified as containing a player are shown in green. *Bottom row*: Player particles are distributed among existing tracks with respect to their likelihoods and posterior track positions are estimated using the associated particles. Estimated position is shown only for the player on the left. Final tracks are shown on the image on the top row and on the soccer field image on the bottom-left.

Chapter 2

Related Work

The research on multiple object tracking is well rooted and applies to a wide range of domains. Reviewing all studies in the tracking literature is beyond the scope of this thesis (See [5, 6, 7, 8, 9, 10] for detailed surveys); thus in this chapter, we give a brief review of the studies most relevant to the domain of sports video analysis. [1]

2.1 Camera Configuration

One of the most important decisions to make when approaching a sports player tracking problem is camera configuration. In [11, 12, 13, 14, 15, 16, 17, 18], broadcast footage captured by a pan-tilt-zoom camera is used, offering a relatively cheap and flexible solution to this issue because it is not necessary to physically set up cameras to track players in a game. However, such approaches must deal with continuous changes in view-point. A more severe problem is that broadcast videos are usually zoomed to the region of action, therefore some players become not visible for tracking. As a solution, some studies ([19, 20, 21, 22, 23]) place a number of static cameras in order to capture a single-view of the entire field.

[1] © 2015 IEEE. This chapter contains text that is reused with permission, from [4].

However, as it can be quite challenging for single-view tracking algorithms to resolve frequent and continuous occlusions of players, the methodologies proposed in [24, 25, 26, 27, 28, 29, 30] tackle the problem by pursuing a multi-view approach, in which the observations from four to eight static cameras are fused. Although the efforts of these multi-view approaches are laudable, considering the structure of sports arenas/stadiums, these systems introduce extra complications such as difficulties in camera setup, the need to route data to a single processing node, and increased computational complexity, which makes them impractical and relatively expensive for real-time applications.

2.2 Player Segmentation

Depending on camera configuration, different approaches have been applied for player segmentation. When using static cameras, the simplest way to segment players on the field is to apply background subtraction or statistical background modeling followed by a set of morphological operations, as in [19, 25, 26, 30]. Background subtraction or modeling is inapplicable if a pan-tilt-zoom camera is being used. Alternatively, assuming color-homogeneity of the field, dominant color analysis on a Hue channel or histogram back-projection can be used to extract a background mask to remove it from the overall image to locate players, as in [11, 14, 17]. In cases of extreme weather or unstable lighting conditions, these simple player segmentation methods would most likely suffer and generate many false positives. Recently, more sophisticated methodologies have been proposed to cope with such conditions. Gedikli *et al.* [12] employ special templates that extract likelihood-maps for player locations based on color distributions, compactness, and vertical spacing cues. Liu *et al.* [15] use a boosted cascade detector using Haar features. Xing *et al.* [18] apply a hybrid multi-cue learning algorithms with online and offline stages. Lu *et al.* [16] utilize a Deformable Part Model to automatically locate players. Given a calibrated camera, player locations are estimated by fitting fixed height 3d cylinders to the foreground mask in [31, 32, 33]. Herrmann *et al.* [13] extracts player confidence maps by applying grass segmentation, and utilizing color and gradient cues.

2.3 Multiple Player Tracking

The problem of tracking multiple sports players has been tackled from different perspectives that can be grouped into three main categories.

2.3.1 Deterministic Methods

Several approaches employ visual features in a deterministic manner to search for a player's track in the next frame. Color templates are used in early approaches, such as [34]. The idea of kernel density estimation, such as the Mean-shift tracker [35], is applied in [14], using color cues. For better tracking performance, shape information can be decoupled from color, as in [36], or texture and local motion vectors can be used in addition to visual color features, as in [21]. Recent methods such as [37], use a kernelized structured output support vector machine, to learn the appearance of the track and adapt to changes. To better represent the target, and to distinguish foreground and background, [38] utilizes a tracking template using discriminative non-orthogonal binary subspace spanned by Haar-like features. Such approaches do not properly encapsulate interactions among players; therefore, these methods are likely to be distracted when players are occluded or similarly colored tracks are near each other.

2.3.2 Data Association and Optimization-based Methods

From another point of view, having detected players in each time unit, one can formulate tracking as a data association problem and seek an optimal solution in a variety of ways. Gedikli *et al.* [12] use a Multiple Hypothesis Tracker [39] to create affiliations between current observations and previous player trajectories. A Joint Probability Data Association Filter [40] is applied to link player observations between consecutive frames in [25, 30]. Figueroa *et al.* [19] construct a graph in such a way that blobs correspond to the nodes, edges represent the distance between the blobs, and players are tracked by traversing the graph by considering the minimal path. Di *et al.* [41], segment blobs in each frame, encode object history into states and describe state transitions through a Finite State Automata (FSA). Shitrit *et al.* [24] formulate a Probabilistic Occupancy Map (POM) of the players as a direct acyclic graph, and find global optimal solution by linear programming. In a follow up study [42], this time POM is utilized by formulating the problem as a Multi-Commodity Network Flow. Lu *et al.* [16] use bipartite matching to associate player detections to existing tracks. Liu *et al.* [33, 32] employ hierarchical data association to track sports players with context-conditioned motion models. These approaches require accurate consecutive observations to correctly establish links and theoretically reach a global optimum. Moreover, they involve explicit detection and exhaustive iteration through all associations in a certain time interval, introducing a heavy computational delay that makes them impractical and rather expensive for real-time applications.

2.3.3 Probabilistic Methods

The Bayesian framework and its estimations offer another solution to the multiple player tracking problem. Random-like movements can be tracked by Sequential Monte Carlo Estimation, also known as Particle Filtering [43], which has recently become a popular tracking methodology due to its ability to cope with uncertainties in visual observations and track non-linear models.

The states of all tracked objects are embodied into a single joint state and particle filtering techniques are applied for tracking in [44]. This approach was also adopted by Czyz *et al.* [45] for tracking soccer players. The problem with the joint-state model is that it has a size bound, therefore only a limited number of players can be tracked; more important, inaccuracies in tracking a single player may affect the entire estimation. Several solutions to this problem have been presented, including Liu *et al.* [15], in which an optimal solution is estimated using a Markov Chain Monte Carlo (MCMC) sampler. Collins *et al.* [31] proposed a hybrid MCMC algorithm that uses deterministic solutions for blocks of variables to accelerate its stochastic mode-seeking behavior.

Another approach to the player tracking problem is to reduce the state-space size and use separate trackers for each player, as in [11, 13, 46, 20, 28, 29, 17, 23]. However, it is crucial for these types of methods to consider players' global state to avoid one player hijacking the track of another due to similar likelihood scores. To cope with this problem, Ok *et al.* [17] use occlusion probability scores; Hess *et al.* [46] present discriminative training methods for tracking American football players that attempt to directly optimize the filter parameters in response to observed errors; Kristan *et al.* [20] take advantage of the bird's-eye camera at indoor sports venues and manage the interactions of individual particle filters using a Voronoi partitioning of the space. Herrmann *et al.* [13] utilize visual evidences such as color and HOG to extract a confidence map and find local maxima to track players. Schlipsing *et al.* [23] employ SVMs to learn appearance of players through color histogram and use a Kalman Filter based multi-object tracking approach.

2.4 Comparison to Related Studies

The particle filtering approaches generate many particles to accurately track each target. Each particle represents a hypothesis for the track, and particles are propagated with respect to an auto-regressive model. Perez *et al.* [47] propose a probabilistic tracker based on particle filters that uses similarity of color histograms for likelihood evaluation. To better handle the multi-modality of the target distribution that may arise due to presence of multiple objects, Vermaak *et al.* [48] extend the work of [47] and introduce a Mixture Particle Filter (MPF), in which each object is modeled with an individual particle filter that forms part of a mixture. In a follow-up study, Okuma *et al.* [49] employ MPF to track hockey players, supported by the Adaboost algorithm [50] for player detection.

The MPF approach performs better than naive particle filtering approaches in resolving basic occlusions among opponents and tracking multiple targets because interactions among the tracks are evaluated by spatially clustering all the particles and allowing particle transfer between different tracks. However, MPF can easily under-perform in soccer videos since teammates look almost identical and players are involved in frequent and continuous occlusions. Such cases result in particle degeneration, in which particles of a track are propagated towards another target or transferred to another mixture component. Hence, identity switches or hijackings may occur among tracks during occlusions.

Instead of employing separate particles for each target on the image plane, we utilize the real-world ground plane, and introduce the idea of densely sampled particles at fixed positions. These particles are spread on a model soccer field such that they represent possible locations for tracks. Multiple targets are probabilistically tracked on these model field particles, in which the likelihood of a track being on a particle is evaluated globally. Our likelihood function that utilizes color, motion and soccer player appearance cues, enables us to properly associate particles with tracks to provide the following advantages over standard particle filtering approaches: Occlusions are handled implicitly resulting in less identity switches and track losses; few particles are needed to accurately track the target resulting in a more efficient tracker; tracking processes is simplified such that there is no need for a particle re-sampling step.

Chapter 3

Our Approach

We present our approach to track multiple sports players. First, we introduce the concept of model field particles (Section 3.1) and provide details on model field construction (Section 3.2). Next, we describe our methodology to detect the player on the model field (Section 3.3) and give details of our proposed multiple player tracking algorithm. Finally, we conclude the chapter by presenting our approach to initialize, recover and correct the identity of the tracks in short and long terms (Section 3.5). [1]

3.1 Concept of Model Field Particles

A soccer field is modeled using a set of densely sampled particles $\mathbf{S} = \{\mathbf{s}^1, \mathbf{s}^2, \mathbf{s}^3, \ldots, \mathbf{s}^{\mathrm{M}}\}$, where M is the total number of particles needed to span the entire field. These particles discretize the possible position of the players on the model soccer field and each particle $\mathbf{s}^m \in \mathbf{S}$ is represented with a quadruple, such that $\mathbf{s}^m = \{\mathbf{q}^m, B^m, \mathbf{a}^m, e^m\}$. The unique two dimensional position of a sampled particle on the model field is denoted with $\mathbf{q}^m$ and each particle is represented by a corresponding bounding box B^m on the image plane, with

[1] © 2015 IEEE. This chapter contains text that is reused with permission, from [4].

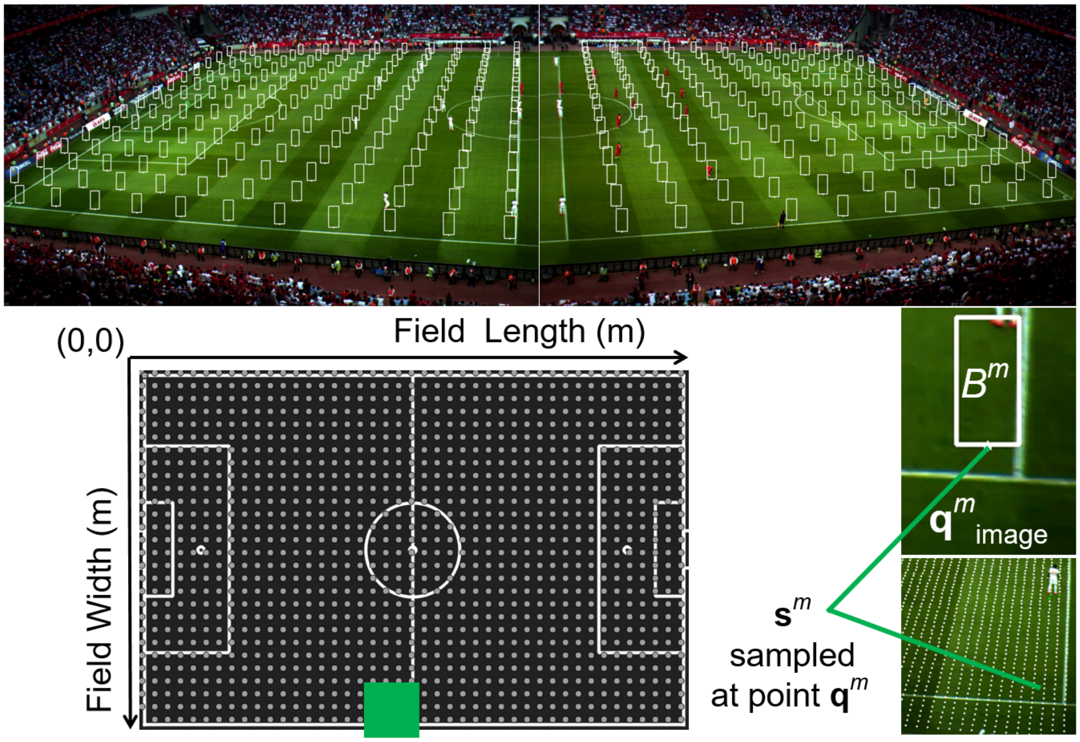


Figure 3.1: A soccer field is modeled by densely sampled particles, $\mathbf{S} = \{\mathbf{s}^1, \mathbf{s}^2, \mathbf{s}^3, \ldots, \mathbf{s}^{\mathrm{M}}\}$, discretizing the possible position of the players. Corresponding bounding boxes for some particles are shown on the image plane. Note that the model field is depicted with sparsely sampled particles for better visualization. In the real case, each square meter contains four particles.

an appearance model $\mathbf{a}^m$, as shown in Figure 3.1. Bounding boxes overlap with each other on the image plane so that a player always employs a set of neighbor particles. A Histogram of Oriented Gradients (HOG) [51] detector, trained for soccer, is used to decide whether $\mathbf{s}^m \in \mathbf{S}$ contains a player by examining its B^m, where e^m is the likelihood of containing a player. It follows that $\mathbf{S}^+ \subset \mathbf{S}$ denotes the subset of positively classified particles that may be occupied by players.

The likelihood of a track being on a particle is evaluated by a combination of appearance and motion models. To grasp the global state of the game and the interactions among players, each particle $\mathbf{s}^m \in \mathbf{S}^+$ is associated with the track having the highest likelihood. In order to cope with occlusions, low probability particles may also be associated with tracks if the motion likelihood of the track is highest for a particle. Finally, tracks are separately propagated using a weighted linear combination of their associated particles.

The color and motion models complement each other in multiple player tracking. Color handles the unpredictable motion patterns since they usually occur when opponents with different colored jerseys are near each other. Motion comes into play when color confuses teammates due to similar appearances. Tactically, teammates show different motion patterns, especially when they are near each other (It is not common for teammates to run side-by-side towards the same direction at same speed). During occlusions, the concept of densely sampled particles and global likelihood calculation with prioritizing motion model enable players to be aware of each other and keep their locations while their view is blocked. To better handle sudden changes in velocity and avoid drifting problem, after distributing particles among the tracks, the final track to particle likelihoods are calculated using only appearance cues.

3.2 Model Field Construction

We densely sample M particles $\mathbf{S} = \{\mathbf{s}^1, \mathbf{s}^2, \ldots, \mathbf{s}^\mathrm{M}\}$ on the model field. Each square meter of the soccer field is spanned by four particles so that a player always stands on many sample particles on the model field. The standard dimensions of the soccer field are 105×68 meters, resulting in M=28,560 particles if a square meter is spanned by 2×2 particles. In the following subsections, we describe our methods for representing model field particles on the image plane.

3.2.1 Camera Configuration

The proposed system uses two high-definition cameras to view the soccer field; one camera is adjusted to capture the left half and the other is adjusted to capture the right half (see Figure 3.2). A narrow portion of the field along the midfield line should be visible in both cameras to establish a homography relation between the tracks in common. The camera synchronization is handled by a software trigger

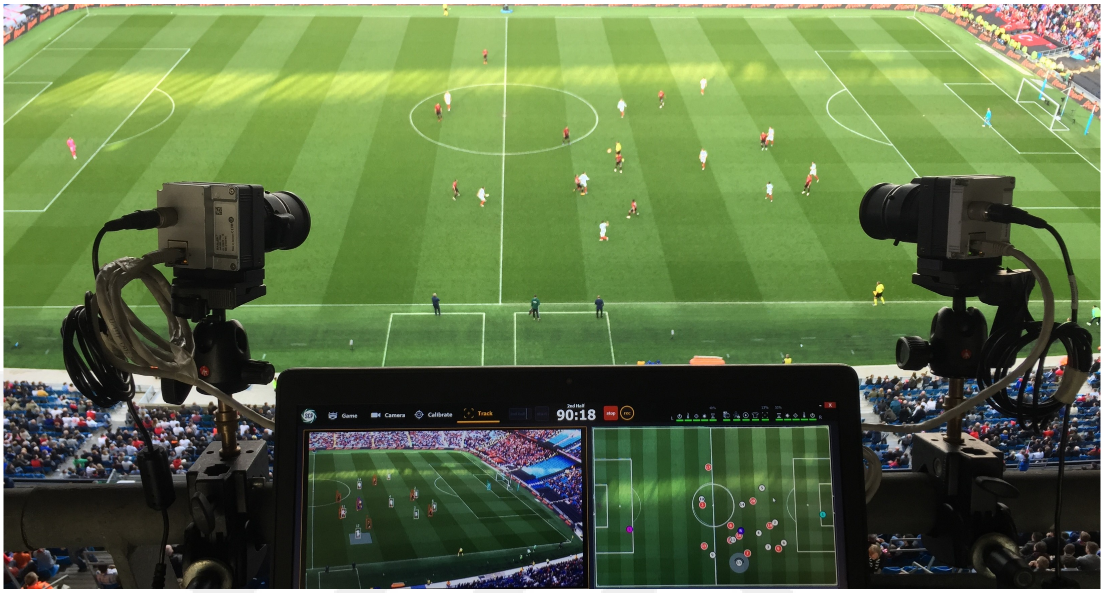


Figure 3.2: Hardware of the proposed system. Images captured from two high-definition cameras are processed on a laptop.

and the exposure is controlled automatically, as in [52], by continuously extracting gray level histograms of the soccer field, excluding the non-field regions in the image, and adjusting the exposure until a target mean gray value is reached. Images acquired from the two cameras are processed on a powerful laptop to execute the proposed multi-player tracking algorithm.

3.2.2 Distortion Elimination

Since the cameras shoot a large area (size of a half is 68×52.5 meters) from close range, the lenses cause radial distortion, resulting in a curved appearance of the actual straight lines in the image. The distortion must be corrected by estimating coefficients, and pixels must be warped to their correct locations. Based on [53], the relation between a distorted point $\mathbf{q}_d = (x_d, y_d)$ and an undistorted point $\mathbf{q}_{image} = (x_{image}, y_{image})$ on the image plane is expressed as

$$\begin{aligned} x_{image} &= c_x + (1 + \kappa_1 r + \kappa_2 r^2 + \kappa_3 r^3 + \ldots)(x_d - c_x), \\ y_{image} &= c_y + (1 + \kappa_1 r + \kappa_2 r^2 + \kappa_3 r^3 + \ldots)(y_d - c_y). \end{aligned} \tag{3.1}$$

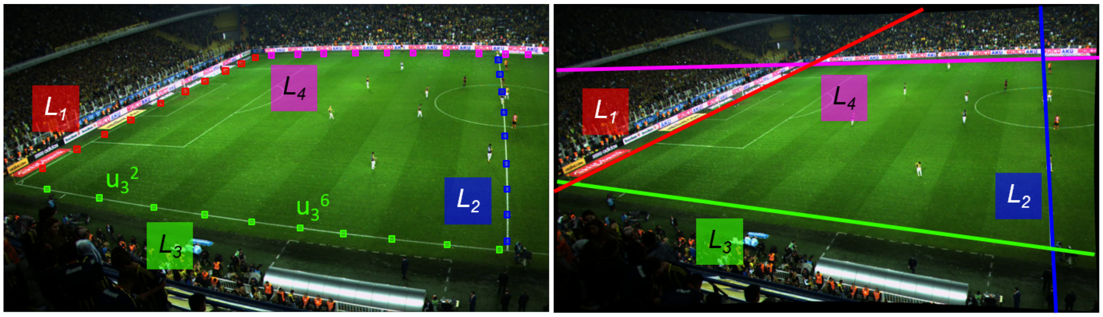


Figure 3.3: *Left*: Calibration points marked on the four field boundary lines in distorted image. Notice the curved appearance of the points on each line. *Right*: Undistorted version of the top image. Observe that the field boundary lines are straight in the undistorted image. L$_1$: goal line, L$_2$: midfield line, L$_3$: near sideline, L$_4$: far sideline. © 2015 IEEE. Reprinted with permission, from [4].

Here, $r^2 = (x_d - c_x)^2 + (y_d - c_y)^2$, $\{\kappa_1, \kappa_2, \kappa_3, \ldots\}$ are the radial distortion correction coefficients, and (c_x, c_y) are the image center coordinates. Note that we only use κ_1 for distortion correction.

The points on the image plane placed on the field boundaries $L_1 \ldots L_4$ are manually marked (see Figure 3.3). These points appear as a curve on the distorted image, but they should form a straight line on the undistorted image. This fact is used to estimate the coefficients by undistorting the marked points using Eq. (3.1) for different values of κ_1, and choosing the value that minimize the average mean squared error when the lines are fitted to the undistorted points as

$$\underset{\kappa_1}{\operatorname{argmin}} \sum_{j=1}^{4} \sum_{\mathbf{u}_j^i \in L_j} \min \|\mathbf{u}_j^i - L_j\|. \tag{3.2}$$

Here, $\mathbf{u}_j^i$ is a point marked on the field boundary line j, and L_j is the line fitted to the corresponding set of points.

3.2.3 Perspective Transformation

Having corrected the image distortion, the perspective transformation between the particle location points on the model field $\mathbf{q}_{model}$ and the points on the undistorted image plane $\mathbf{q}_{image}$ are defined as $\mathbf{q}_{image} = H \cdot \mathbf{q}_{model}$ (Note that in the following we refer to $\mathbf{q}_{model}$ as $\mathbf{q}$ for simplicity.)

Given a set of at least four point correspondences, the homography matrix H can be estimated using Direct Linear Transformation [54]. We use the four corners of the soccer field in the image plane (which are extracted by intersecting the field boundary lines $L_1 \dots L_4$) and their correspondences on the model field. Note that more point correspondences can be used to reduce the calibration error. Separate homography matrices H_{left} and H_{right} are calculated for the left and right cameras. Particles on the left and right halves of the model soccer field correspond to left camera (points are transformed using H_{left}) and right camera (points are transformed using H_{right}) images, respectively.

3.2.4 Particle Representation on the Image Plane

Each model field particle $\mathbf{s}^m = \{\mathbf{q}^m, B^m, \mathbf{a}^m, e^m\}$ is described by its position $\mathbf{q}^m = (x, y)$ and its appearance $\mathbf{a}^m$ obtained from the corresponding bounding box B^m, on the image plane. Consider a player standing on a particle $\mathbf{s}^m$ at position $\mathbf{q}^m$. The corresponding point on the image plane $\mathbf{q}^m_{image} = H\mathbf{q}^m$ is approximated using perspective transformation, as described in the previous subsection. Then, the height of B^m, which should be tall and wide enough to encapsulate a player, is estimated in pixels to correspond to a fixed height T_{player}, in meters on $\mathbf{q}^m_{image}$.

The rule of perspectivity states that parallel lines intersect at a vanishing point. As observed in Figure 3.4, the line that connects the two vanishing points of the border line pairs (L_1, L_2) and (L_3, L_4) is the horizon. Since the soccer field is planar, all the imaginary perpendicular lines drawn from the horizon to the soccer field ground in the image plane actually have the same height in the

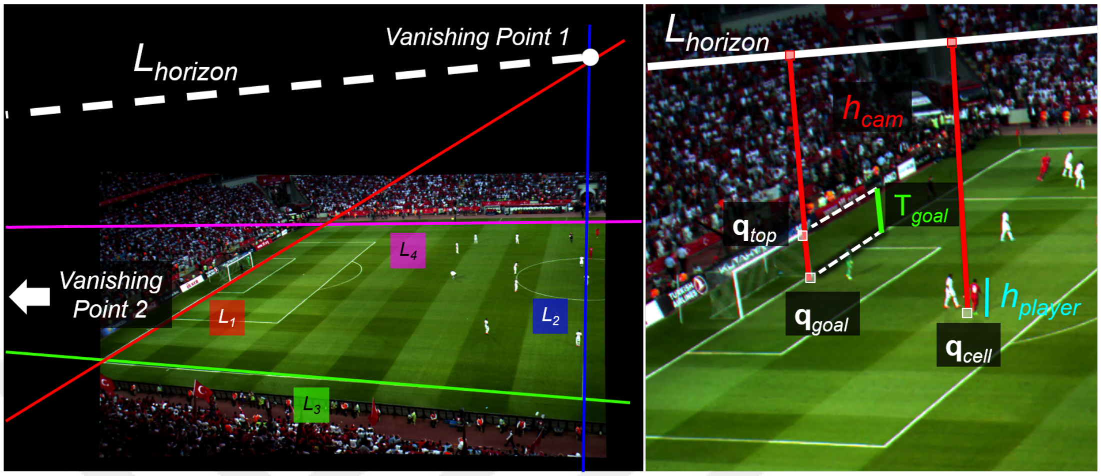


Figure 3.4: Illustrates the representation of model field particles on the image plane with bounding boxes. *Left*: Using field boundary lines to find vanishing points and the horizon line. *Right*: The calculation of a bounding box height h_{player} (in pixels) corresponding to a target height T_{player} (in meters). A reference object with a known height in meters T_{goal} is utilized to derive the camera height h_{cam}. Then the camera height and the distance to the horizon are used to calculate the height of each bounding box B^m in the image.

real world, corresponding to the height of the camera above the ground. This principle is utilized to calculate a fixed-height (in meters) bounding box for each model field particle, whereas the bounding box heights in pixels can be different due to the perspective effect. Using the goal posts as reference objects, with a known height of 2.44 meters, the bounding box height in pixels for each particle is calculated using direct proportion as

$$\begin{aligned} h_{cam} &= \left(\min\|\mathbf{u}_{bottom} - L_{horizon}\| \cdot \mathrm{T}_{goal}\right) / \|\mathbf{u}_{bottom} - \mathbf{u}_{top}\|, \\ h_{player} &= \left(\min\|\mathbf{q}^m_{image} - L_{horizon}\| \cdot \mathrm{T}_{player}\right) / h_{cam}. \end{aligned} \tag{3.3}$$

As visualized in Figure 3.4, here $L_{horizon}$ is the horizon line, $\mathbf{u}_{bottom}$ and $\mathbf{u}_{top}$ are the bottom and top of the goal post in the image, T_{goal} is the fixed height of the goal post (equal to 2.44 meters) and h_{cam} is the camera height in meters, T_{player} is a fixed constant for the target bounding box height in meters, and h_{player} is the height of the bounding box (in pixels) to be calculated at $\mathbf{q}^m_{image}$. The target height T_{player} is set to 1.90 meters and the width of B^m is set to half of its height.

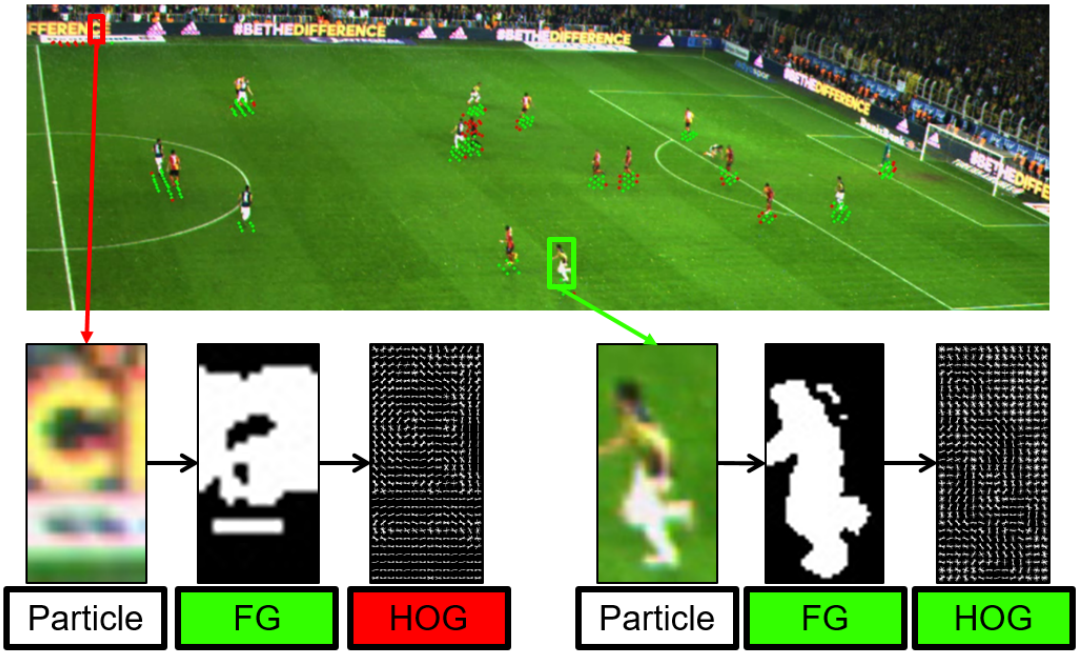


Figure 3.5: Steps of player detection. Each particle having a ratio of foreground pixels above some threshold are considered as a candidate to contain a player, then f-HOG [55] features are extracted for these particles, and finally an SVM classifier is used to decide if the particle contains a player. *Top*: Candidate particles that are classified as positive (green) and negative (red). *Bottom*: Two candidate particles that are classified as negative and positive, respectively. Raw image, foreground image and f-HOG vector illustration is shown from left to right.

3.3 Player Detection

It is far from reality to expect the standard background subtraction-based approaches to leave only the pixels belonging to the players. In matches played under sunlight or in the absence of sufficient illumination, simple shadow detection algorithms are likely to fail at eliminating dark player shadows on the field. Moreover, pixels belonging to the same player may be broken into separate blobs, or a single blob may contain pixels belonging to more than one player. We utilize the concept of model field particles and propose an approach for locating players on the soccer field that is robust to challenging illumination conditions (see Figure 3.5 for illustration of the approach).

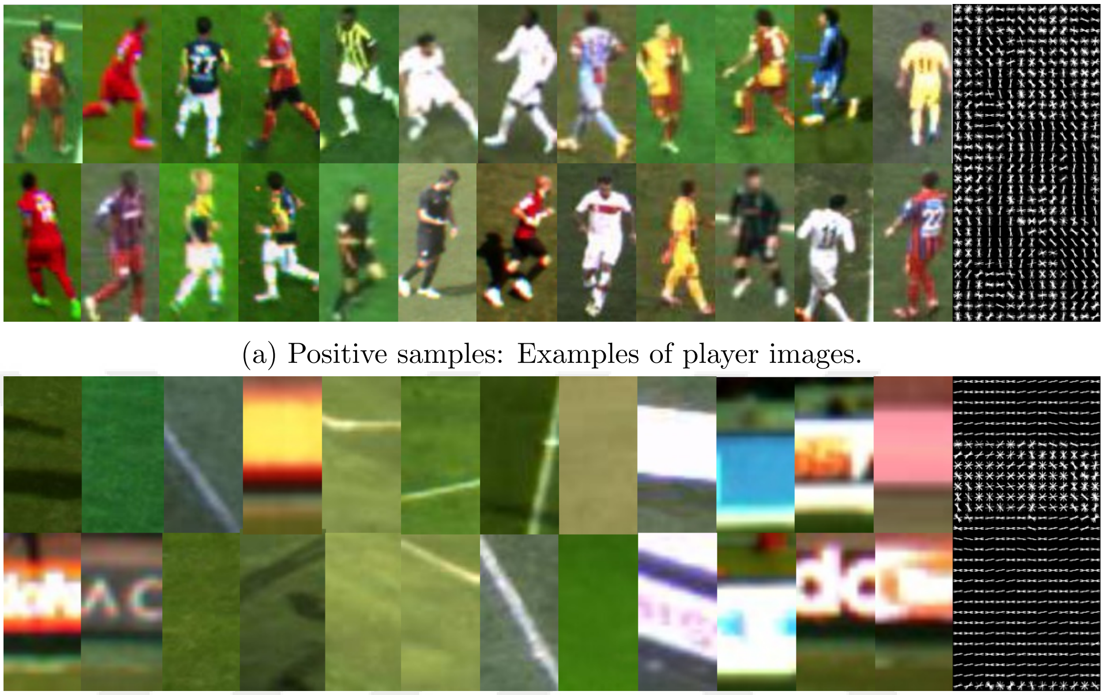
3.3.1 Foreground Extraction

First, we exploit the foreground segmentation to reduce the number of candidate regions for players. Given an image, the foreground is extracted using the adaptive Gaussian mixture model described in [56, 57]. Then median filtering and morphological closing operation are applied for noise removal. Alternative to the fixed global learning rate, we propose using a dynamic spatial learning rate, which is more suitable for soccer videos. The learning rate is automatically adjusted to reconstruct the mixture model if a sudden increase in the number of foreground pixels is detected (indicating a rapid change in lighting). In addition, the learning rates of digital billboard pixels are set to relatively higher values for quick adaptation to continuously changing and blinking ads.

3.3.2 Supervised Player Classification

We aim to decide if a particle $\mathbf{s}^m \in \mathbf{S}$ is occupied by a player. Since the large number of particles is difficult to exhaustively traverse and process even if it is done in parallel, we reduce the number of model field particles to be examined by extracting the foreground regions, as described in the previous subsection. However, during sudden light changes or in presence of dark player shadows, a lot of false positive foreground pixels will be generated. To ignore the particles with falsely extracted foreground regions, we utilize a classifier for player detection.

We employ a HOG-based [51] method for human detection due to its abilities to efficiently describe complex shapes and edges in different scales, tolerate small deformations and cope with illumination and contrast variances. Recall that each sample particle $\mathbf{s}^m \in \mathbf{S}$ on our model field has a corresponding bounding box B^m as a potential image patch that may encapsulate a player. Each bounding box B^m is divided into three spatial regions vertically and if all the regions have a ratio of foreground pixels above some threshold, then $\mathbf{s}^m$ is considered as a candidate to contain a player. The threshold for foreground ratio is empirically set to 15%.



(a) Positive samples: Examples of player images.

(b) Negative samples: Examples of non-player images.

Figure 3.6: Illustrates sample images among 120,000 used for training the soccer player classifier. Black and white images on the right depict f-HOG [55] features for a positive and a negative sample, respectively. Notice how the f-HOG illustration of the negative sample contains homogeneously oriented gradients that makes it distinguishable from f-HOG vectors of positive samples.

For each $\mathbf{s}^m \in \mathbf{S}'$, where $\mathbf{S}' \subset \mathbf{S}$ is the subset of candidate particles to contain a player, the image patch $I[B^m]$ described by the bounding box B^m is resized to a constant $height \times width$ pixels, divided into overlapping spatial cells and a 31-dimensional f-HOG [55] vector is extracted for each cell. These f-HOG vectors are concatenated and normalized to obtain the final descriptor for $\mathbf{s}^m \in \mathbf{S}'$. The f-HOG descriptors are classified by a linear Support Vector Machine (SVM) [58] classifier hSVM, trained using a wide spectrum of 60,000 player and 60,000 non-player samples collected from over 20 soccer videos with different environmental conditions (see Figure 3.6 for examples of positive and negative samples).

Classification scores of hSVM model are transformed into a probability distribution over classes using Platt scaling [59], which works by fitting a logistic regression model to the scores. It follows that each candidate particle $\mathbf{s}^m \in \mathbf{S}$ has a player detection likelihood e^m that is calculated as

$$e^m = \frac{1}{1 + \exp(-\mathrm{hSVM}(hog(I[B^m])))}, \tag{3.4}$$

where hog uses the image patch $I[B^m]$ described by B^m to extract f-HOG vector and hSVM uses this f-HOG vector to return a classification score. Only the set of positively classified model field particles $\mathbf{S}^+ \subset \mathbf{S}$ is used in tracking the players. In the following, for simplicity we refer to $\mathbf{s}^m$ as a model field particle that is positively classified, and discard the particles that are negatively classified. That is, we will only consider $\mathbf{s}^m \in \mathbf{S}^+$. Note that, since the operations applied to each sample particle are exactly the same, we distribute the process of player detection to multiple processors.

3.3.3 Track Initiation

As observed in Figure 3.5, a player stands on many neighbor model field particles with overlapping bounding boxes on the image plane. To initiate a new track, the overlapping detections are merged by using the idea of non-maximum suppression [60]. A new track is created at the location of the particle with the local maximum player detection likelihood in Eq. (3.4). The neighbor particles having overlapping bounding boxes with the local maximum particle are ignored. Two bounding boxes are said to be overlapping if their geometric centers are closer than some threshold distance. For merging detections along the midfield line, we use plane-to-plane homography to transform geometric centers between images for those bounding boxes that are distributed across different cameras. Only those particles that are not occupied by existing players are used for new track initiation.

3.4 Multiple Player Tracking

3.4.1 Problem Formulation

A sports match can be represented by a collection of consecutive states and their forward transitions. The state of the game at any instant can be described using a set of features encapsulating the players' positions, their visual appearances, motion models, and interactions. Then, the objective of tracking multiple players is to estimate the state of the game $\mathbf{x}_t$ at time t, given a set of observations $\mathbf{z}_{1:t}$ up to the present time. If this is assumed to be a first-order Markov process, denoted as $p(\mathbf{x}_t|\mathbf{z}_{1:t})$, then the posterior estimation can be characterized in two steps; the first involving the prediction of the next state from prior knowledge, and the second performing an update with new observation data [43]:

$$p(\mathbf{x}_t|\mathbf{z}_{1:t-1}) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1})\ p(\mathbf{x}_{t-1}|\mathbf{z}_{1:t-1})\ d\mathbf{x}_t, \tag{3.5}$$

$$p(\mathbf{x}_t|\mathbf{z}_{1:t}) \propto p(\mathbf{z}_t|\mathbf{x}_t)\ p(\mathbf{x}_t|\mathbf{z}_{1:t-1}). \tag{3.6}$$

As implied by the prediction (Eq. (3.5)) and update (Eq. (3.6)) equations, the posterior estimation process requires specifying the state-space dynamics for describing the state evolution $p(\mathbf{x}_t|\mathbf{x}_{t-1})$, as well as the existence of a model that evaluates the likelihood of an observation for a given state $p(\mathbf{z}_t|\mathbf{x}_t)$. We present an efficient and effective estimation of the stochastic process, in which each player is represented with a disjoint state and tracked separately. The game's global dynamics and player interactions are captured through the observation model by employing the model field particles as measurements of the states and distributing them among the tracks using a combined appearance and motion likelihood model.

3.4.2 State-Space Dynamics

The state of the game at time t can be defined as the collection of individual player states $\mathbf{X}_t = \{\mathbf{x}_t^1, \mathbf{x}_t^2, \ldots, \mathbf{x}_t^{\mathrm{N}}\}$, where N is the total number of players/tracks. The state of a player/track $\mathbf{x}_t^n \in \mathbf{X}_t$ is defined as

$$\mathbf{x}_t^n = \begin{bmatrix} \mathbf{p}_t^n & \mathbf{v}_t^n & \mathbf{b}^n \end{bmatrix}, \tag{3.7}$$

where $\mathbf{p}_t^n = (x, y)$ is the two-dimensional position of the player on the model soccer field, $\mathbf{v}_t^n$ is the velocity, and $\mathbf{b}^n$ is the reference appearance model of the target being tracked.

3.4.2.1 State Prediction

Omitting the appearance $\mathbf{b}^n$, a Kalman Filter [61] with a constant velocity motion model is used for handling each player state $\mathbf{x}_t^n \in \mathbf{X}_t$, and the prediction of the next state is made as

$$p(\mathbf{x}_t^n | \mathbf{x}_{t-1}^n) \propto F\ \mathbf{x}_{t-1}^n + \omega_t, \tag{3.8}$$

where $F = [1\ \Delta t;\ 0\ 1]$ is the state transition model, $\omega_t \sim N(0, Q)$ is the process noise representing acceleration (which is assumed to be drawn from a zero mean multivariate normal distribution with covariance $Q = [\frac{\Delta t^4}{4}\ \frac{\Delta t^3}{2};\ \frac{\Delta t^3}{2}\ \Delta t^2]\ \sigma_{acc}^2$ is the acceleration variance, and Δt is the time between two states expressed in seconds (which is set to 1 / frames per second (FPS)). In the following, all the calculations are described for a single time instant t.

3.4.2.2 State Update

Recall that the model soccer field $\mathbf{S}$ is spanned by densely sampled particles and a player detector extracts the subset $\mathbf{S}^+ \subset \mathbf{S}$ of particles that denote the possible locations of the tracks on the model field. Each particle $\mathbf{s}^m \in \mathbf{S}^+$ can be represented with the quadruple $\mathbf{s}^m = \{\mathbf{q}^m, B^m, \mathbf{a}^m, e^m\}$. Here, $\mathbf{q}^m = (x, y)$ is the fixed two-dimensional location of $\mathbf{s}^m$ on the model soccer field, B^m is the pre-calculated bounding box on the image plane, $\mathbf{a}^m$ is the current appearance model of the image patch described by B^m, and e^m is the likelihood of the particle to contain a player.

At each time instant t, the model field particles are distributed among the players with respect to the likelihood of track $\mathbf{x}^n$ being on $\mathbf{s}^m$, denoted as $p(\mathbf{s}^m|\mathbf{x}^n)$, which is calculated using a combined appearance and motion model. Then the final measurement $\mathbf{p}^n_{observed} = (x, y)$ of track $\mathbf{x}^n$, indicating the observed position at time t, presumed to be corrupted by a noise ϵ, is calculated as

$$\mathbf{p}^n_{observed} \sim \sum_{\mathbf{s}^m \in f(\mathbf{x}^n)} w(\mathbf{x}^n, \mathbf{s}^m) \cdot \mathbf{q}^m + \epsilon, \tag{3.9}$$

where $f : \mathbf{X}_t \to \mathbf{S}^+$ is a functional relation and $f(\mathbf{x}^n) \subset \mathbf{S}^+$ is the subset of particles associated with track $\mathbf{x}^n$; $w(\mathbf{x}^n, \mathbf{s}^m)$ is the weight of $\mathbf{s}^m$ extracted by normalizing the likelihood values of track $\mathbf{x}^n$ being on all of the associated particles such that each $w(\mathbf{x}^n, \mathbf{s}^m) \propto p(\mathbf{s}^m|\mathbf{x}^n)$ and $\sum_{\mathbf{s}^m \in f(\mathbf{x}^n)} w(\mathbf{x}^n, \mathbf{s}^m) = 1$; and $\epsilon \sim N(0, Z)$ is the measurement noise, assumed to be a zero mean Gaussian white noise with covariance Z, which is chosen so that the maximum error is approximately the shoulder width (0.5 meters).

Then, given the observation $\mathbf{p}^n_{observed}$ for track $\mathbf{x}^n$, a state update is made using the standard Kalman Filter update equations in [61]. The reference appearance model $\mathbf{b}^n$ of the track is extracted from the image patch corresponding to the player's position on the model field and is updated every second by a weighted addition to cope with pose and illumination changes. The appearance model of a track is updated only when its set of particles are not neighbors with the particles assigned to any other track.

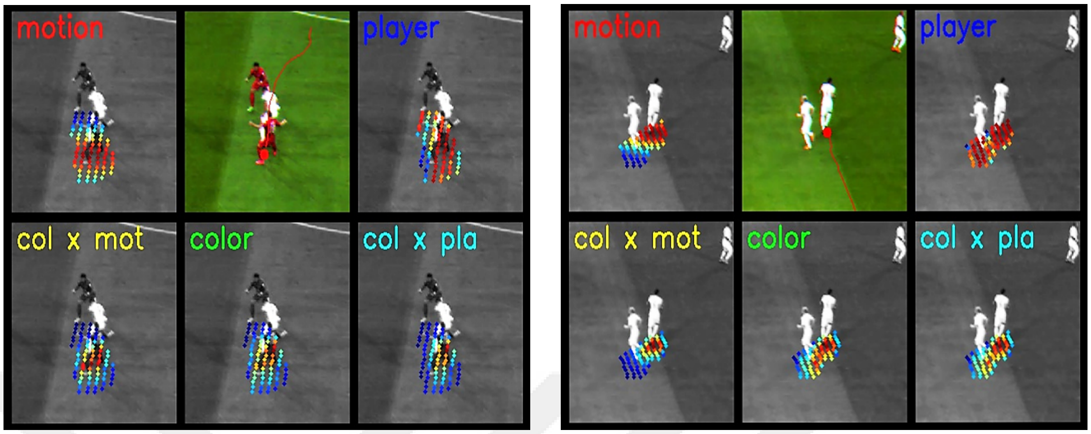
3.4.3 Likelihood Models

The nature of soccer requires players to be spatially separated as much as possible from the teammates, and as close as possible to their opponents since they are involved in possession challenges and tackles. Therefore, color is an important cue to capture the diversity in the appearance of opponents wearing different jerseys. However, utilizing only color features may result in identity hijackings and tracking ambiguities among nearby teammates. As a solution, we propose coupling color features with the target's motion model which yields better tracking of players with similar appearances.

The likelihood of a track $\mathbf{x}^n \in \mathbf{X}_t$ being on a particle $\mathbf{s}^m \in \mathbf{S}^+$ at a time t is evaluated separately for appearance and motion models; then these independent probabilities are multiplied to obtain the overall likelihood. Then, particles are distributed among the tracks with respect to their overall likelihood. After the particle distribution, track positions are estimated by a weighted combination of the associated particles, where weight of a particle is in proportion with its likelihood. However, color features are coupled with player detection scores ($e^m \in \mathbf{s}^m$) instead of the target's motion model. As it will be shown in the experimental results, after the particle distribution, using color likelihood and player detection scores together better captures the non-linearity in target's motion compared to using color and motion features (see Figure 3.7 for visualization of different likelihood models).

3.4.3.1 Appearance Model

The employed appearance model should be able to handle illumination effects and capture the spatial layout of the color distribution on the players' jerseys. The methods proposed in [47, 49] are able to successfully cope with such problems.



(a) Tracking a player near opponents

(b) Tracking a player near a teammate

Figure 3.7: Likelihood of a player being on the densely sampled model field particles. Top center image shows the tracked target with a red dot; and images around the top center illustrates likelihood of the player being on the particles with respect to different likelihood models. Values are normalized and visualized in a jet color map, in which blue and red represents the lowest and highest probabilities respectively.

Following these studies, we extract an appearance model $\mathbf{a}^m$ for each $\mathbf{s}^m \in \mathbf{S}^+$ by dividing B^m into upper and lower regions and formulating Hue-Saturation-Value (HSV) histograms for each spatial region. An HSV histogram $\mathbf{a}$ is composed of a concatenation of separate HS and V channel histograms, with a total of $\mathrm{C} = \mathrm{C}_h\mathrm{C}_s + \mathrm{C}_v$ bins and $\mathbf{a}[c]$ denotes the number of pixels in the c-th bin, where $c \in \{1, 2, \dots, \mathrm{C}\}$ is the bin index. Each histogram $\mathbf{a}$ is normalized to represent the color model as a probability distribution such that $\sum_{c=1}^{C} \mathbf{a}[c] = 1$. The reference histogram $\mathbf{b}^n$ of each track $\mathbf{x}^n \in \mathbf{X}_t$ is calculated in the same way as the model field particles.

To calculate the color likelihood $p_{color}(\mathbf{s}^m|\mathbf{x}^n)$, the reference color histogram $\mathbf{b}^n$ of track $\mathbf{x}^n$ is compared to the histogram of particle $\mathbf{a}^m$ using the Bhattacharyya similarity coefficient. It follows that distance d_{color} between the two color histograms is defined as

$$d_{color}(\mathbf{b}^n, \mathbf{a}^m) = \left(1 - \sum_{c=1}^{C} \sqrt{\mathbf{b}^n[c]\ \mathbf{a}^m[c]}\right)^{1/2} . \qquad (3.10)$$

It is reported in [47] that successful tracking runs based on color similarity yield consistent exponential behavior for the squared distance d^2_{color}; thus the color likelihood of a track being on a particle is defined as

$$p_{color}(\mathbf{s}^m|\mathbf{x}^n) \propto \exp -\lambda \frac{1}{J} \sum_{j=1}^{J} d^2_{color}\big(\mathbf{b}^n_j, \mathbf{a}^m_j\big), \tag{3.11}$$

where $J = 2$ is the number of subregions (upper and lower body), and $\mathbf{b}^n_j$ and $\mathbf{a}^m_j$ are the color histograms extracted from the j-th subregion of the image patches belonging to $\mathbf{x}^n$ and $\mathbf{s}^m$ respectively. In our experiments, we achieved the best results when the number of bins in the HSV histogram was set to $\mathrm{C}_h = 10$ and $\mathrm{C}_s = \mathrm{C}_v = 5$ when $\lambda = 20$, as in [47].

3.4.3.2 Motion Model

Recall that positional information is maintained by a Kalman Filter and the posterior state of the track is predicted using Eq. (3.8), based on prior knowledge. The motion model evaluates the likelihood of a track $p_{motion}(\mathbf{s}^m|\mathbf{x}^n)$ by simply measuring the distance d_{motion} between the predicted position of the track and the location of the particle on the model soccer field such that

$$d_{motion}\big(\mathbf{p}^n, \mathbf{q}^m\big) = \|\mathbf{p}^n - \mathbf{q}^m\|. \tag{3.12}$$

Here, $\mathbf{p}^n$ is the predicted position of track $\mathbf{x}^n$ and $\mathbf{q}^m$ is the location of $\mathbf{s}^m$. The motion likelihood is inversely proportional to d_{motion} since it is higher for the particles closer to the predicted position and decreases as the distance between the predicted position and the particle location increases. As a result, the motion likelihood of a track being on a particle, which can be modeled as a normal distribution around the predicted position, is defined using a delta function δ such that

$$\delta(d) = \frac{1}{\sigma_{motion}\sqrt{\pi}} \exp - \frac{d^2}{\sigma_{motion}^2}, \tag{3.13}$$

$$p_{motion}(\mathbf{s}^m|\mathbf{x}^n) \propto \delta\big(d_{motion}(\mathbf{p}^n, \mathbf{q}^m)\big). \tag{3.14}$$

Here, σ_{motion} is the standard deviation of the normal distribution determining the interval of the motion likelihood values. Recall the bell shape of a normal distribution and note that choosing a relatively low σ_{motion} will result in a more pointy curve and hence, a larger penalty will be applied as the distance between the predicted position and the particle location increases.

3.4.3.3 Combined Appearance and Motion Model

A combined color and motion model is used for calculating an overall likelihood to distribute the particles among the tracks at each time instant. If follows that, the likelihood of a track $\mathbf{x}^n \in \mathbf{X}_t$ is evaluated separately by the appearance and motion models, using Eq. (3.11) and Eq. (3.14) respectively. Then the overall likelihood is calculated by multiplying the independent probabilities such that

$$p_{color \times motion}(\mathbf{s}^m|\mathbf{x}^n) \propto p_{color}(\mathbf{s}^m|\mathbf{x}^n) \cdot p_{motion}(\mathbf{s}^m|\mathbf{x}^n). \tag{3.15}$$

Observe on Figure 3.7b that motion balances color in the probability multiplication, in order to avoid high likelihood (due to color similarity) between tracks and particles that are far away from each other. This range is controlled by σ_{motion}, which determines the process noise of the motion model, and acts as the impact factor of motion on the overall likelihood. A lower value of σ_{motion} will result in dramatically decreasing motion likelihood values as the distance between the predicted position and the particle location increases. In contrast, a higher σ_{motion}, narrow the scale of motion likelihood values, and hence increase the impact of color in Eq. (3.15).

3.4.3.4 Appearance Model with Player Detection Score

At each time instant, the appearance model is combined with player detection scores for weighting the particles associated with each track to estimate final track position. Then, this estimated position is used as a observation to update the track state by Eq. (3.9). If follows that, the likelihood of a track $\mathbf{x}^n \in \mathbf{X}_t$ being on a particle, is evaluated by the appearance model using Eq. (3.11) and multiplied with the player detection score of the particle to get the overall likelihood in Eq. (3.16).

$$p_{color \times player}(\mathbf{s}^m|\mathbf{x}^n) \propto p_{color}(\mathbf{s}^m|\mathbf{x}^n) \cdot e^m, \tag{3.16}$$

where the player detection score for a particle is constant for all tracks.

3.4.4 Global Likelihood Evaluation

There are many instances in soccer in which the individual player trackers with different likelihood models may fail. These occasions include opponents being completely occluded during tackles, teammates standing still near each other so that their similar appearance may result in identity switches, and a bunch of interacting players in challenge of possession during set pieces. To resolve tracking ambiguities in such cases, players' spatial locations with respect to each other should be utilized and the game's global state must be encapsulated in the tracking algorithm. Hence, we propose to (the process of the global likelihood evaluation is depicted in Figure 3.8):

i. Distribute the model field particles among the tracks at each instant with respect to their likelihoods. A particle is assigned to the track having the highest combined color and motion likelihood.

ii. Estimate the position of a track by the weighted combination of its assigned

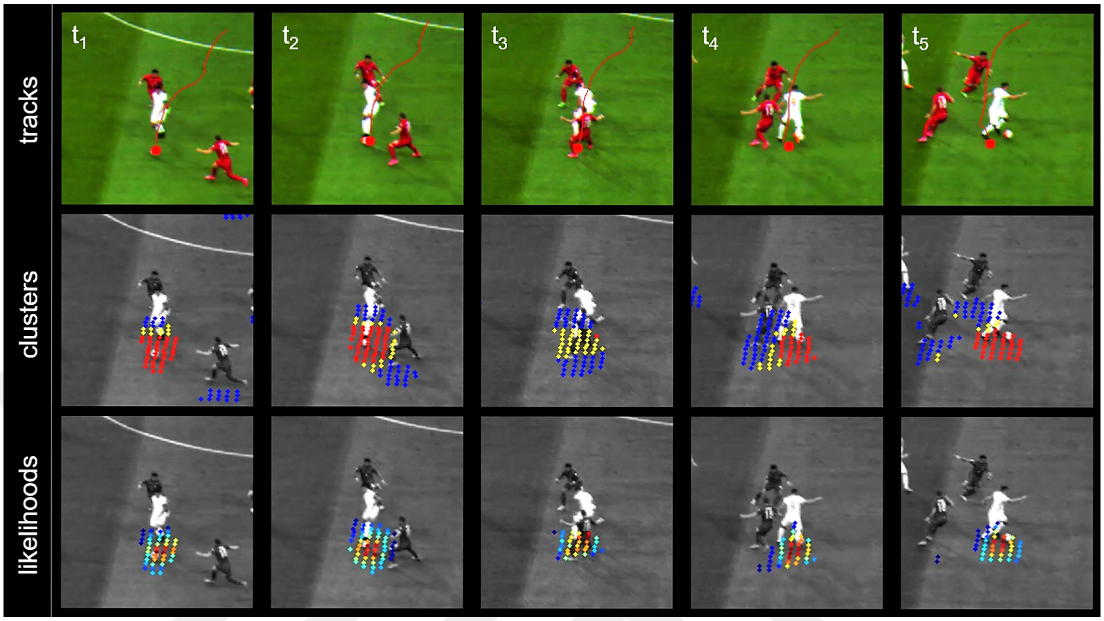


Figure 3.8: (Best seen in color) Tracking a player through occlusion by global likelihood evaluation and distributing particles among tracks. Each column represents a time instant, where $t_1 < t_2 < \ldots < t_5$ and t_1 is the oldest. *Top row*: Red dot shows the estimated position of the tracked player and red line shows the prior path. *Middle row*: Distribution of particles among the players. Red particles belong to the tracked player, blue particles belong to the others, and yellow particles are shared. *Bottom row*: Weighting of the particles assigned to the tracked player. Weights are normalized and visualized in a jet color map, in which blue and red represents the lowest and highest probabilities respectively.

particles, in which color likelihood is combined with player detection scores for weighting.

iii. Allow particles to be shared among the tracks in order to handle occlusions. Independent of its overall likelihood during particle distribution, a track keeps a particle if it has the highest motion likelihood.

3.4.4.1 Particle Distribution Among Tracks

At each time instant t, we define a functional relation $g : \mathbf{S}^{+} \rightarrow \mathbf{X}_t$ where $g(\mathbf{s}^m) \subset \mathbf{X}_t$ denotes the subset of tracks claiming to be on the particle $\mathbf{s}^m$. Each track $\mathbf{x}^n$ claims to be on all nearby particles $\mathbf{s}^m$ such that $\|\mathbf{p}^n - \mathbf{q}^m\| < r_{max}$, where r_{max} is the search radius around the predicted position $\mathbf{p}^n$ of track large enough to include the particles that the player can travel in Δt. Each particle $\mathbf{s}^m$ is assigned to the occupying track $\mathbf{x}^n \in g(\mathbf{s}^m)$ having the highest likelihood $p_{color \times motion}(\mathbf{s}^m|\mathbf{x}^n)$. Then the observation for a track $\mathbf{x}^n$ is obtained as in Eq. (3.9) by assigning a weight $w(\mathbf{x}^n, \mathbf{s}^m)$ to each associated particle $f(\mathbf{x}^n) \subset \mathbf{S}^+$, where $f : \mathbf{X}_t \to \mathbf{S}^+$ is a functional relation from tracks to the set of model field particles, $w(\mathbf{x}^n, \mathbf{s}^m) \propto p_{color \times player}(\mathbf{s}^m|\mathbf{x}^n)$ and $\sum_{\mathbf{s}^m \in f(\mathbf{x}^n)} w(\mathbf{x}^n, \mathbf{s}^m) = 1$.

3.4.4.2 Occlusion Handling by Motion Model

A player who is partially or completely occluded by an opponent may have low likelihoods on all nearby particles and hence be lost because none of the particles will be associated with the track. An occlusion can only occur on the image plane and since motion model is evaluated on the real-world ground plane, it is utilized in tracking players through occlusions. A track $\mathbf{x}^n \in \mathbf{X}_t$ is said to be *occluded* on a particle $\mathbf{s}^m \in \mathbf{S}^+$, if it has a lower overall likelihood compared to the other tracks but has the highest motion likelihood on the particle, such that

$$\exists \mathbf{x}^i \in \mathbf{X}_t \ (\mathbf{x}^i \neq \mathbf{x}^n) : p_{color \times motion}(\mathbf{s}^m|\mathbf{x}^n) < p_{color \times motion}(\mathbf{s}^m|\mathbf{x}^i), \tag{3.17}$$

$$\forall \mathbf{x}^i \in \mathbf{X}_t \ (\mathbf{x}^i \neq \mathbf{x}^n) : p_{motion}(\mathbf{s}^m|\mathbf{x}^n) > p_{motion}(\mathbf{s}^m|\mathbf{x}^i). \tag{3.18}$$

It follows that if a track $\mathbf{x}^n$ is *occluded* on a particle $\mathbf{s}^m$, the assignment $f(\mathbf{x}^n) \leftarrow \mathbf{s}^m$ is preserved independent of the other tracks claiming to be on $\mathbf{s}^m$. In other words, particles are assigned to tracks having the highest overall or motion likelihood. One might think that the algorithm may fail to cope with non-linearity in motion since the motion model overrides the overall likelihood. However, this is not the case because, the track positions are calculated without motion model, using $p_{color \times player}(\mathbf{s}^m|\mathbf{x}^n)$ after the particle distribution. The complete multi-player tracking methodology is formalized in Algorithm 1

Algorithm 1: Iteration of our multi-player tracking methodology at time t. $p_{c\times m}$ denote $p_{color\times motion}$ and $p_{c\times p}$ denote $p_{color\times player}$.

```
Data: Set of model field particles $\mathbf{S}^+$ at $t$
Result:  Update state of each track $\mathbf{x}^n \in \mathbf{X}_t$

/* Predict next state and calculate the likelihoods                  */
foreach $\mathbf{s}^m \in \mathbf{S}^+$ do $g(\mathbf{s}^m) \leftarrow \varnothing$;
foreach $\mathbf{x}^n \in \mathbf{X}_t$ do $f(\mathbf{x}^n) \leftarrow \varnothing$;
foreach $\mathbf{x}^n = \begin{bmatrix}\mathbf{p}^n & \mathbf{v}^n & \mathbf{b}^n\end{bmatrix} \in \mathbf{X}_t$ do
    $p(\mathbf{x}^n_t|\mathbf{x}^n_{t-1}) \propto F_t\ \mathbf{x}^n_{t-1} + \omega_t$ ;
    foreach $\mathbf{s}^m = \left\{\mathbf{q}^m, B^m, \mathbf{a}^m, e^m\right\} \in \mathbf{S}^+$ do
        if $\|\mathbf{p}^n - \mathbf{q}^m\| < r_{max}$ then
            $p_{c\times m}(\mathbf{s}^m|\mathbf{x}^n) \propto p_{color}(\mathbf{s}^m|\mathbf{x}^n) \cdot p_{motion}(\mathbf{s}^m|\mathbf{x}^n)$ ;
            $p_{c\times p}(\mathbf{s}^m|\mathbf{x}^n) \propto p_{color}(\mathbf{s}^m|\mathbf{x}^n) \cdot e^m$ ;
            $f(\mathbf{x}^n) \leftarrow \mathbf{s}^m$;
            $g(\mathbf{s}^m) \leftarrow \mathbf{x}^n$;
        end
    end
end

/* Globally evaluate likelihoods and distribute particles            */
foreach $\mathbf{s}^m \in \mathbf{S}^+$ do
    foreach $\mathbf{x}^n \in g(\mathbf{s}^m)$ do
        // If not the maximum color $\times$ motion likelihood
        if $\exists \mathbf{x}^i \in \mathbf{X}_t\ (\mathbf{x}^i \neq \mathbf{x}^n)$ : $p_{c\times m}(\mathbf{s}^m|\mathbf{x}^n) < p_{c\times m}(\mathbf{s}^m|\mathbf{x}^i)$ then
            // If not the maximum motion likelihood
            if $\exists \mathbf{x}^j \in \mathbf{X}_t\ (\mathbf{x}^j \neq \mathbf{x}^n)$ : $p_{motion}(\mathbf{s}^m|\mathbf{x}^n) < p_{motion}(\mathbf{s}^m|\mathbf{x}^j)$ then
                $f(\mathbf{x}^n) = f(\mathbf{x}^n) - \{\mathbf{s}^m\}$;
            end
        end
    end
end

/* Update track positions using associated particles                 */
foreach  $\mathbf{x}^n \in \mathbf{X}_t$ do
    $w(\mathbf{x}^n, \mathbf{s}^m) \propto p_{c\times p}(\mathbf{s}^m|\mathbf{x}^n)$ and $\sum_{\mathbf{s}^m \in f(\mathbf{x}^n)} w(\mathbf{x}^n, \mathbf{s}^m) = 1$ ;
    $\mathbf{p}^n_{observed} \sim \sum_{\mathbf{s}^m \in f(\mathbf{x}^n)} w(\mathbf{x}^n, \mathbf{s}^m) \cdot \mathbf{q}^m + \epsilon$;
end
```

3.5 Player Identification

To extract any kind of player-specific data, the identities of the tracks corresponding to the real players must be known. Initially, identity tags are manually assigned to the tracks just before the kick-off. Then, we employ a regional collective motion model and an optimal assignment-based algorithm to recover from track losses in the short-term. However, some of the tracks may not be assigned an identity tag in the short-term or identity of the tracks may be switched during occlusions. To cope with such situations, a positional appearance learning model is employed to maintain the correct player identities in the long-term.

3.5.1 Jersey Classification

Based on their jersey colors, tracks in a soccer match can be classified as belonging to one of five teams or classes: home/away team goalkeeper, home/away team player, and referee. During the manual assignment of the track identities before kick-off, clusters representing each of the jersey classes are initiated by retrieving sample color histograms from each class. The reference histogram R^n of a newly created track $\mathbf{x}^n \in \mathbf{X}_t$ is compared to the samples in each cluster, using the color similarity function in Eq. (3.11), and team identity is assigned using k-nearest-neighbor (k-NN) classification. To capture the diversity in player appearance due to pose changes and varied illumination in different regions of the field, and to increase the classification accuracy, we maintain clusters with many jersey samples and automatically update clusters at regular intervals throughout the game.

Jersey classification is the initial step of all of the player identification procedures. Assigning the correct jersey class to a newly created track is sufficient for maintaining goalkeepers' and the referee' identity tags because they have distinctive jerseys and spatial regions of action on the field. Classifying a new track as a home or an away player by jersey color narrows down the solution space to half size. The problem then becomes predicting the home/away player identity to be assigned to the track. The following sections describe complementary methodologies for assigning home/away player identity tags to the tracks in short and long terms.

3.5.2 Collective Motion Model

Depending on the state of the game, observe on Figure 3.9 that closely located group of soccer players tend to move collectively in the same direction, possibly with similar speeds. For instance, at short intervals of time, two opponents may be chasing for the ball or defenders may be holding their positions as a block to defend their zone. In the following, we describe the utilization of this game context feature to seek for the lost players by evaluating the collective motion around them.

Excluding the goalkeepers, the identity tags of home/away team players are assigned as follows. Let $Y_k = \{y_k^1, y_k^2, \dots, y_k^{10}\}$ be the set of real player identities (denoting the jersey number), where $k \in \{1, 2\}$ indicates the home and away team and let $\mathbf{X}_k \subset \mathbf{X}$ (the time notation is dropped for readability) be the set tracks having the team identity k assigned by jersey classification. Then, mapping the set of unassigned player tags $Y'_k \subset Y_k$ to the set of tracks with no player identities $\mathbf{X}'_k \subset \mathbf{X}_k$ can be formulated as an optimal assignment problem and solved using the Hungarian method [62]. The cost of assigning $y_k^i \in Y'_k$ to $\mathbf{x}_k^n \in \mathbf{X}'_k$ is denoted and set as $cost\big(\mathbf{x}_k^n \leftarrow y_k^i\big) = \|\mathbf{p}_k^n - \mathbf{pp}_k^i\|$, where $\mathbf{p}_k^n$ is the current track position and $\mathbf{pp}_k^i$ is the estimated position of the unassigned player. While a player identity tag $y_k^i \in \mathbf{Y}_k$ is assigned to a track $\mathbf{x}_k^n \in \mathbf{X}_k$, denoted by $tag(\mathbf{x}_k^n) \leftarrow y_k^i$, the estimated position of the player is continuously updated by the track such that $\mathbf{pp}_k^i = \mathbf{p}_k^n$.

When a track is lost, its corresponding player identity $\mathbf{y}_k^i \in \mathbf{Y}'_k$ is marked as lost as well. The estimated position $\mathbf{pp}_k^i$ of the player, points to the last seen location of the associated track at the time of loss, and $\mathbf{pp}_k^i$ is updated at each time instant by a weighted combination of the positions of the nearest tracks in $\mathbf{X}$, as follows. Let $\{\mathbf{r}^1, \mathbf{r}^2, \dots, \mathbf{r}^{\mathrm{J}}\}$ be the set of sorted 2d track positions in ascending order with respect to their closeness to $\mathbf{pp}_k^i$, then $\mathbf{pp}_k^i = \sum_{j=1}^{\mathrm{J}} w(n)\, \mathbf{r}^j$, where $w(n) \propto 1/\|\mathbf{pp}_k^i - \mathbf{r}^j\|$ is the weight of the track and $\sum_{j=1}^{\mathrm{J}} w(n) = 1$. Note that a track $\mathbf{x}^n \in \mathbf{X}$ is considered in the neighborhood of a point $\mathbf{pp}_k^i$, if the distance between track position and point is smaller than a search radius such that $\|\mathbf{pp}_k^i - \mathbf{p}^n\| < r_{search}$. The search radius r_{search} is initiated with a small value and incremented in each time instant with respect to the distance that the lost player can travel in that time interval.

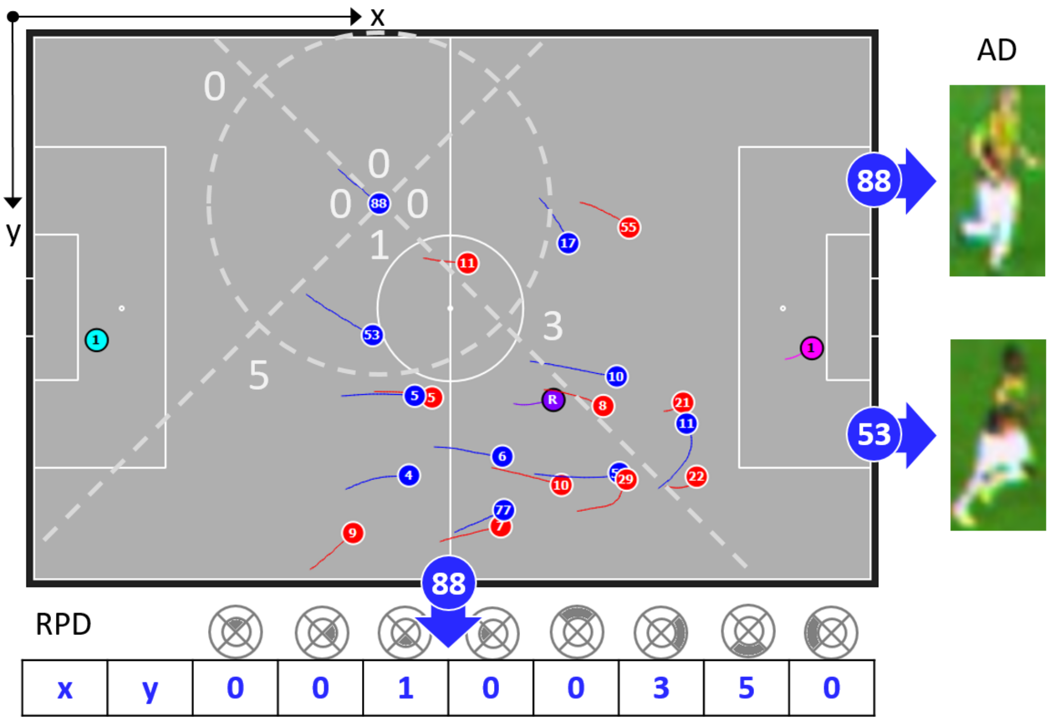
The Hungarian algorithm is run throughout the game whenever $|\mathbf{Y}'_k| > 0$ and $|\mathbf{X}'_k| \geq |\mathbf{Y}'_k|$ to assign lost player identities to the new tracks. Since the game state and player positions may change significantly, the short-term search for a lost player identity is terminated after 20 seconds if an assignment could not be made.

3.5.3 Positional Appearance Learning

Just like most of the team sports, soccer is also played in pursuit of a team tactic or strategy in which players occupy different areas of the field depending on their duty and generally keep their relative positioning with respect to the other players. We utilize the idea of relative occupancy of the players and combine it with appearance cues to learn and distinguish the identity of the players throughout the game. For instance, by positional appearance learning, we aim to distinguish a blond left-back, a tall central midfielder, and a striker wearing orange shoes.

3.5.3.1 Relative Position Descriptor

Inspired by the shape context descriptor in [63] and its application to team sports as Relative Occupancy Maps in [33], similarly, we describe relative position of a track by a distance-orientation histogram. Referring to the notation from Sec 3.5.2, a track $\mathbf{x}_k^n \in \mathbf{X}_k$ is compared to the other tracks $\mathbf{x}_k^i \in \mathbf{X}_k$ having the same team identity to calculate distance $d(\mathbf{x}_k^n, \mathbf{x}_k^i) = \|\mathbf{p}_k^n - \mathbf{p}_k^i\|$ and orientation $\theta(\mathbf{x}_k^n, \mathbf{x}_k^i) = \arctan(\mathbf{p}_k^n, \mathbf{p}_k^i) \cdot 180/\pi$.



x	y	0	0	1	0	0	3	5	0

Figure 3.9: Process of extracting Relative Position Descriptor (RPD) and Appearance Descriptor (AD) for the track with number 88. Soccer field is partitioned into 8 regions around the track with respect to orientation and distance. An RPD is calculated by counting the number of teammates (in blues) in each region to form a histogram, where first two bins denote the normalized position of the track on the field. ADs (to be normalized and converted into a row vector) are extracted by resizing image patches to a constant size, so that tracks can be distinguished using the raw pixels.

The Relative Position Descriptor (RPD) of $\mathbf{x}_k^n$, denoted by $rpd(\mathbf{x}_k^n)$, consists of 10 bins. The first two bins describe the normalized track location by the soccer field dimensions. Remaining bins represent the count of the teammates, centered around the track, in different distance and orientation regions. The orientation space is divided into four main orientations and the distance space is divided into near and far regions by a distance threshold r_{pos}. Each (d, θ) pair contributes to the representing histogram bin and the histogram is finalized by applying L2-normalization, excluding the first and second bins. Figure 3.9 illustrates the process of extracting an RPD for a track.

3.5.3.2 Appearance Descriptor

Unlike RPDs, an appearance descriptor (AD) is extracted from the image plane and occlusions may occur on the image plane. Therefore, an AD is extracted in each time instant for those tracks that have no other tracks nearby, which indicates a clear, line-of-sight view of the track. Since players on the image can be considered as low-resolution, raw color pixels are directly used for describing the appearance of a track $\mathbf{x}_k^n \in \mathbf{X}_k$. Observe on Figure 3.9 that two different players can be distinguished by utilizing cues like skin and shoe color. The appearance descriptor for $\mathbf{x}_k^n$, denoted by $ad(\mathbf{x}_k^n)$, is extracted from the image patch $I[B^j]$ by resizing it to a constant $height \times width$ pixels, converting it to a row vector, and applying min-max normalization. Here, B^n is the corresponding bounding box (on the image) to the model field point $\mathbf{p}^n$ and it is calculated as described in Sec. 3.2.4.

3.5.3.3 Tag Tracks by Positional Appearance Model

Different SVM classifiers pSVM$_k$ and aSVM$_k$ are trained for home (k=1) and away (k=2) teams with RPDs and ADs, respectively. The feature sampling begins with the manual initiation of the tracks just before the kick-off. If a track is tagged such that $tag(\mathbf{x}_k^n) = y_k^i$ and $\mathbf{x}_k^n$ is in line-of-sight, $rpd(\mathbf{x}_k^n)$ and $ad(\mathbf{x}_k^n)$ are extracted and added to the set of sample descriptors with the label y_k^i. Once there are enough samples for each player from a team, pSVM$_k$ and aSVM$_k$ are trained. The sampling continues throughout the game and classifiers are re-trained in every 10 minutes using the last 1000 samples for each player. Especially at the beginning of the match when there are not enough samples, augmented samples are generated from existing descriptors by randomly rotating and scaling image patches for ADs, and adding random noise to player locations for RPDs.

Throughout the game, each track $\mathbf{x}_k^n \in \mathbf{X}_k$ is classified by both pSVM$_k$ and aSVM$_k$, whenever $\mathbf{x}_k^n$ is in line-of-sight at time t. If the classification labels agree such that $y_k^i = \text{pSVM}_k(rpd(\mathbf{x}_k^n)) = \text{aSVM}_k(ad(\mathbf{x}_k^n))$, the possibility of tagging $\mathbf{x}_k^n$

as y_k^i at time t, is added to the global set of classifications $\Phi \leftarrow \{t, \mathbf{x}_k^n, y_k^i\}$. This global set Φ is kept up to date by discarding tags that are older than 20 seconds. When a classification is made, it is not directly applied to track $\mathbf{x}_k^n$. The set Φ is exploited to derive global probability $p_{tag}(y_k^i|\mathbf{x}_k^n)$ of assigning identity y_k^i to track $\mathbf{x}_k^n$. If Φ contains at least η classifications of $\mathbf{x}_k^n$ and $p_{tag}(y_k^i|\mathbf{x}_k^n) > \tau$, then $tag(\mathbf{x}_k^n) \leftarrow y_k^i$ is applied to tag or correct the identity on track $\mathbf{x_k}^n$. Probability threshold τ must be at least set to 0.5; however, we empirically found that 0.75 is a safer threshold not to corrupt the correct tags on the tracks. Finally, at least $\eta = 15$ classifications are needed to apply a tag to a track. Positional appearance classification is summarized in Algorithm 2.

Algorithm 2: Iteration of positional appearance classification at time t

```
Data: Set of tracks $\mathbf{x}_k^n \in \mathbf{X}_k$ for teams $k$ = 1, 2; and global set of
      classifications $\Phi$
Result: Tag/correct identity of track $\mathbf{x}_k^n \in \mathbf{X}_k$
foreach $k \in [1 \quad 2]$ do
    /* Discard classifications that are older than 20 seconds */
    foreach $\phi = \{t_o, \mathbf{x}_k^n, y_k^i\} \in \Phi$ do
        if $t - t_o > 20\ seconds$ then
            $\Phi = \Phi - \phi$ ;
        end
    end
    /* Classify tracks and tag/correct their identities        */
    foreach $\mathbf{x}_k^n \in \mathbf{X}_k$ do
        if $isIntersecting(\mathbf{x}_k^n, \mathbf{x}^j)$ = false : $\forall \mathbf{x}^j \in \mathbf{X}$ then
            $y_k^i = \mathrm{aSVM}_k(ad(\mathbf{x}_k^n))$ ;
            if $y_k^i = pSVM_k(rpd(\mathbf{x}_k^n))$ then
                $\Phi \leftarrow \{t, \mathbf{x}_k^n, y_k^i\}$ ;
                if $tag(\mathbf{x}_k^n) \neq y_k^i$ and $size(\mathbf{x}_k^n \in \Phi) > \eta$ and $p_{tag}(y_k^i|\mathbf{x}_k^n) > \tau$
                 then
                    $tag(\mathbf{x}_k^n) \leftarrow y_k^i$ ;
                end
            end
        end
    end
end
```

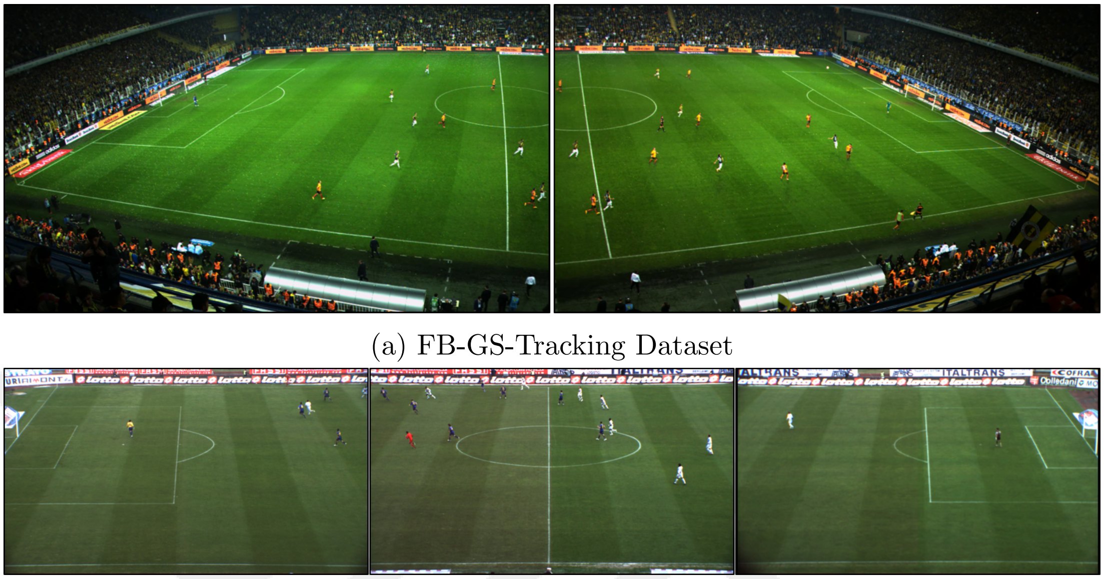
Chapter 4

Experiments

We present the experimental results to evaluate our approach. First, we introduce the datasets (Section 4.1) and criteria used to evaluate tracking (Section 4.2). Later, we measure the accuracy of player detection (Section 4.3), jersey/team classification (Section 4.4), and player identity classification (Section 4.5). Then, we evaluate the proposed multiple player tracking algorithm (Section 4.6) and analyze its computational cost. Finally, we compare our approach with the state-of-the-art single-object trackers (Section 4.7.1) and with other multi-player tracking methods that report their results on publicly available datasets (Section 4.7.2).

4.1 Datasets

i. **FB-GS-Tracking Dataset.** We captured the beginning of the second half of the Turkish Super League soccer match played between Istanbul rivals Fenerbahçe and Galatasaray on 25 October 2015 using two full-HD cameras, as described in Section 3.2.1. Both videos are 10 minutes long and are captured at 10 frames per second (FPS), resulting in 6,000 frames per camera. All 22 players were manually annotated on the calibrated videos to extract



(a) FB-GS-Tracking Dataset

(b) ISSIA-Tracking Dataset [64]

Figure 4.1: Multi-player tracking datasets used in the experimental evaluation.

their ground truth positions and image patches in each frame. Sample frames are illustrated in Figure 4.1a.

ii. **ISSIA-Tracking Dataset.** The publicly available ISSIA dataset [64] consists of 3,000 frames captured by six cameras at 25 FPS, placed around a stadium in a multi-view configuration. The dataset provides two-minutes of game play from a soccer match including ground truth tracking data of all the players. Since our multi-player tracking methodology is currently implemented for single-view, only three cameras in a single-view configuration were used in the experiments. Sample frames are displayed in Figure 4.1b.

iii. **Sparsely-Annotated-Tracking Dataset.** This larger-scale dataset is composed of 15 full-length 90-minute soccer matches collected from the 2015-2016 season of the Turkish Super League. Each game was played in a different stadium in which the two cameras (configured as described in Section 3.2.1) were placed at different heights above the ground. Instead of annotating all the players in each frame, players were annotated in each 20 seconds to extract their ground truth positions. This creates 270 tags for each player in a match. It follows that a match consists of 5,940 tags, summing up to 89,100 tags for the whole dataset.

iv. **Player Dataset.** This binary classification dataset consists of image patches that are gathered from over 20 soccer match videos, in which the games were played at different times of the day and under various environmental conditions. Images patches were cropped from a calibrated camera and as explained in Section 3.2.4, they all correspond to the same fixed height in the real-world. Dataset contains 60,000 positive and 60,000 negative sample images for players and non-players, respectively.

v. **Jersey Dataset.** This is a multi-class dataset that is sub-sampled from the *Player Dataset*. It contains 15 classes representing distinct teams with different jersey colors. Each team/class consists of 100 images of players wearing the same jersey in various poses.

4.2 Evaluation Criteria of Tracking

Multiple Object Tracking Accuracy (MOTA) [65] has become a standard metric for evaluating multiple object trackers and is defined as

$$\text{MOTA} = 1 - \frac{\sum_t \left(c_f(fp_t) + c_m(fn_t) + c_s(mme_t)\right)}{\sum_t G_t}, \tag{4.1}$$

where G_t is the number of ground truth objects in the t-th frame, fp_t is the number of false positives, fn_t is the number of false negatives, mme_t is the number of instantaneous identity switches. Most of the studies follow [66] and set the weighting functions as $c_m = c_f = 1$ and $c_s = \log_{10}$. MOTA is not suitable for evaluating sports player tracking, where identity preserving is crucial, since an identity switch is penalized by term mme_t only on the frame that it occurs. MOTA is dominated by the number of false positives and false negatives which actually measures Multiple Object Detection Accuracy (MODA). Hence, we only use MOTA to compare our approach to the related studies.

MOTA is modified in [24, 42] to define Global Multiple Object Tracking Accuracy (GMOTA), which replaces mme_t with a new term $gmme_t$ that penalizes identity switches globally. Initially, player identity tags are manually assigned to the tracks. Then, in each frame, fn_t is incremented for each missing player; fp_t is incremented for each track without an identity; and $gmme_t$ is incremented for each player whose identity label contradicts with the ground truth identity; both m_t and fp_t are incremented for those tracks whose distance between ground truth position and observation is more than some threshold (1m in our results). We employ the following components of GMOTA to measure the accuracy of our methodology:

$$\text{FP} = \frac{\sum_t fp_t}{\sum_t G_t}, \quad \text{FN} = \frac{\sum_t fn_t}{\sum_t G_t}, \quad \text{GMME} = \frac{\sum_t gmme_t}{\sum_t G_t}. \tag{4.2}$$

Consider two trackers. The first tracker makes 10 mistakes on the last few frames, whereas the second tracker makes a single mistake on the first few frames. Mistakes result in identity switches and they are penalized by the evaluation metrics. Although the first tracker makes a lot more mistakes, it would have a very high GMOTA compared to the second tracker (since $gmme_t$ will be incremented throughout the whole sequence for the second tracker) and a negligible MOTA. To better evaluate the ability of our tracker and to measure the effect of identity switches, we define Local Identity Preserving Accuracy (LIPA) which divides the ordered ground truth data into 20-second intervals, re-initializes track identities at the beginning of each interval and counts the identity miss-matches at the end of each interval. LIPA is defined as

$$\text{LIPA} = 1 - \frac{\sum_{\Delta t \in \mathbf{T}} mme_{\Delta t}}{\sum_{\Delta t \in \mathbf{T}} G_{\Delta t}}, \tag{4.3}$$

where $\mathbf{T}$ is the set of ordered local intervals, Δt is a local interval in $\mathbf{T}$, $mme_{\Delta t}$ is the number of identity switches that occur in Δt, and $G_{\Delta t}$ is the number of targets to be tracked in Δt. Note that the *FB-GS-Tracking Dataset* is divided into $\mathbf{T} = 30$ local intervals of 20 seconds each, and there are a total number of

	HOG Features		f-HOG Features	
Image Size	**Dimension**	**Accuracy**	**Dimension**	**Accuracy**
32 × 16	432	95.37%	372	97.40%
40 × 20	864	96.73%	744	98.09%
48 × 24	1440	96.82%	1240	**98.14**%

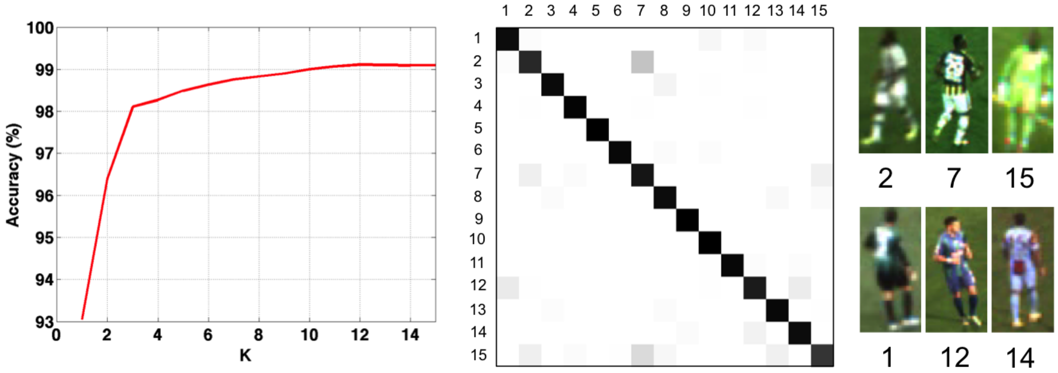
Table 4.1: Accuracy of the SVM classifier on player detection. Each row represents a different configuration in which samples are scaled to different image sizes (*height* × *width* pixels). The classification accuracy with respect to the image size, and the resulting vector dimension is given on each row for different features extractors: HOG [51] and f-HOG [55].

660 tracklets (divisor of fraction in Eq. (4.3)) to measure the accuracy.

4.3 Accuracy of Player Detection

We use *Player Dataset* to evaluate the accuracy of the player detection methodology presented in Section 3.3. We apply ten-fold cross-validation in which 90% of the dataset is used for training the linear SVM player classifier and the remaining samples are used for testing. The effect of scaling image patches to different constant sizes are examined using both the standard HOG features [51] and the f-HOG features [55] employed in our methodology. Features are extracted using a cell size of 4×4 pixels and larger contrast normalization blocks of 2×2 cells. SVM is trained with $c = 1$.

Observe in Table 4.1 that a high binary classification accuracy (**98.14%**) is achieved when the bounding box B^m of each particle $\mathbf{s}^m \in \mathbf{S}$ is scaled to a constant $height \times width$ of 48 × 24. Using f-HOG features instead of the standard HOG features, result in a higher classification accuracy by using smaller feature vectors. The high-accuracy player detector allows us to cope with challenging illumination conditions and successfully locate players on the soccer field. Note that the classification errors can also be tolerated by our tracking methodology during run time. False alarms generated at an unrealistic distance away from the existing players are eliminated by the tracker. Moreover, some of the false alarms



(a) (b) (c)

Figure 4.2: (a) Accuracy of jersey/team classification with respect to k when k-NN leave-one-out cross-validation is applied. (b) Confusion matrix of jersey classification when $k = 1$ (accuracy is 93%). (c) The top and the bottom rows show the set of most-confused team classes in the confusion matrix because of similar jerseys. © 2015 IEEE. Reprinted with permission, from [4].

that are generated close to the existing tracks would possibly get low color and motion likelihood scores and hence can not disrupt the tracks.

4.4 Accuracy of Jersey/Team Classification

The performance of jersey/team classification is crucial in assigning correct identity tags to the tracks and maintaining them throughout the game, as explained in Section 3.5. Jersey assignment of a track also reflects the accuracy of the base color likelihood function since the assignment is simply made by comparing the color histogram with the references using Eq. (3.11).

Jersey Dataset is used in the experiments. Fig. 4.2a shows the classification accuracy graph for different values of k when k-NN leave-one-out cross-validation is applied. Using $k = 10$ yields a jersey/team classification performance of **99%**, which enables us to successfully distinguish the teams in run time and construct the basis for accurately maintaining player identities. Although classification performance is relatively low when $k = 1$, observe in the confusion matrix in Fig. 4.2b that the majority of the errors are made on jerseys with very similar

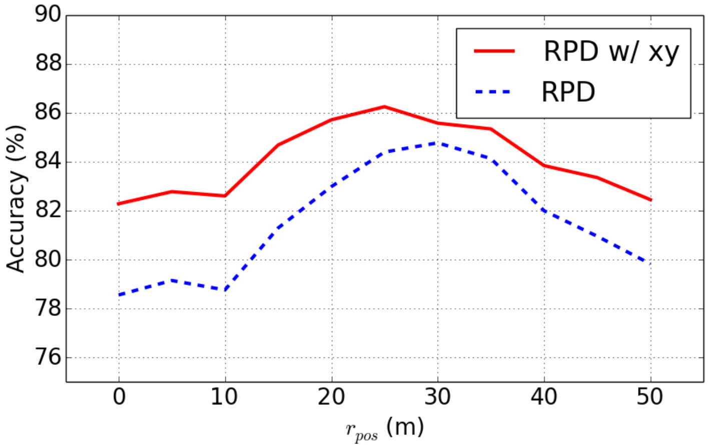


Figure 4.3: Accuracy of relative position classification with respect to the distance threshold r_{pos}. Highest accuracy of **86.2%** is achieved at $r_{pos} = 25$ meters when the descriptors (RPD) are constructed with (w/) normalized xy-locations.

appearance. Note that teams would most likely not wear these similar jerseys when playing with each other.

4.5 Accuracy of Identity Classification

Before measuring the effect of player identification (described in Section 3.5) in multi-player tracking, first, we evaluate the performance of Relative Position Descriptor (RPD) and Appearance Descriptor (AD) in distinguishing the identity of the players after jersey classification. In the experiments, image patches and positions with player identity labels are used from the *FB-GS-Tracking Dataset*. Excluding the goalkeepers, there are 10 players in each team with 6000 samples per player. Ten-fold cross-validation is applied in which 90% of the identity set is used from training linear SVMs (with $c = 0.1$) and the remaining samples are used for testing. Separate pSVM$_k$ and aSVM$_k$ is trained for home ($k = 1$) and away ($k = 2$) teams using RPDs and ADs, respectively.

Figure 4.3 plots the changes in the accuracy of the relative position classifier pSVM$_k$ (averaged for $k = 1, 2$) with respect to the distance threshold r_{pos} (in

Feature Type		Image Size	Accuracy
Raw Pixels	Grayscale Image	32 × 16	62.2%
		40 × 20	59.0%
		48 × 24	55.8%
	Colored Image	32 × 16	**88.2**%
		40 × 20	87.1%
		48 × 24	86.5%
f-HOG [55]	Grayscale Image	32 × 16	64.6%
		40 × 20	74.1%
		48 × 24	78.6%

Table 4.2: Accuracy of appearance classification with respect to different feature types and image patch sizes.

meters). We evaluate constructing RPDs with and without the normalized xy-locations of the players; and seek for the optimal threshold to divide regions into near and far during histogram construction. Results show that, including normalized location of a track in the relative position histogram improves the accuracy of identity classification. RPDs enable pSVM$_k$ to reach a classification accuracy of $\mathbf{86.2}\%$ at $r_{pos} = 25$ meters.

Table 4.2 shows the accuracy of the appearance classifier aSVM$_k$ (averaged for $k = 1, 2$) when different feature types and image patch sizes are used in training. Scaling raw pixels of player images to smaller sized patches better distinguish the appearances since the disrupting details are washed out. The remaining pixels contain cues about player height, skin and color of shoes. Unlike raw pixels, f-HOG features require larger image patches to better classify player identities. The highest accuracy of **88.2%** is achieved when ADs are constructed by using raw pixels, scaling colored image patches to a constant *height* × *width* of 32 × 16 pixels, normalizing the image, and converting it into a row vector.

In positional appearance learning, a classification is considered as confident and applied when both pSVM$_k$ and aSVM$_k$ agree on the label of a track. As shown on Table 4.3, relative position and appearance classifiers provide the same identity label in 72.2% of the cases in which a high accuracy of **96.4%** is achieved.

Classifier	**Accuracy**	**Intersection**
Relative Position	86.2%	N/A
Appearance	88.2%	N/A
Positional Appearance	96.4%	72.2%

Table 4.3: Accuracies of relative position (pSVM$_k$) and appearance (aSVM$_k$) classifiers. In positional appearance learning, classification is made if both pSVM$_k$ and aSVM$_k$ provide the same label (intersection).

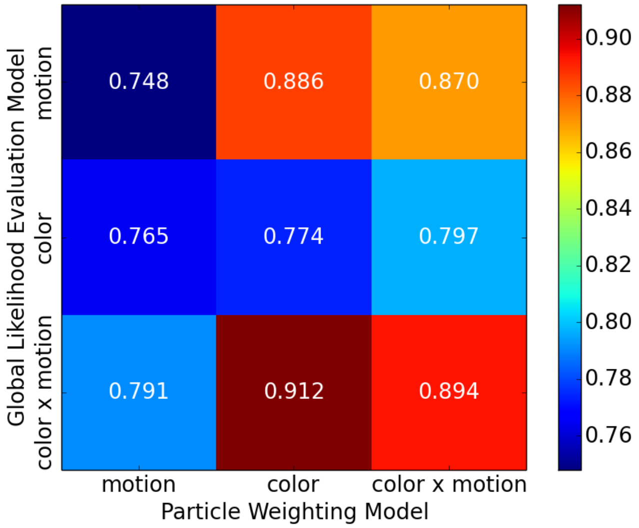


Figure 4.4: Local Identity Preserving Accuracy (LIPA) on *FB-GS-Tracking Dataset* with respect to the model used in global likelihood evaluation to distribute particles among tracks and in weighting particles associated with each track. Motion sigma $\sigma_{motion} = 2$ in all color × motion likelihood calculations.

4.6 Evaluation of Multiple Player Tracking

4.6.1 Evaluation of Likelihood Model

Recall from Section 3.4.3 that color and motion model is used to calculate the likelihood for distributing particles among tracks; color likelihood and player detection score is used for weighting the assigned particles. Figure 4.4 shows LIPA on *FB-GS-Tracking Dataset* for different models used in global likelihood evaluation to distribute particles and in particle weighting for each track after particle-to-track association. Occlusion Handling (`OH`) by motion model (described in

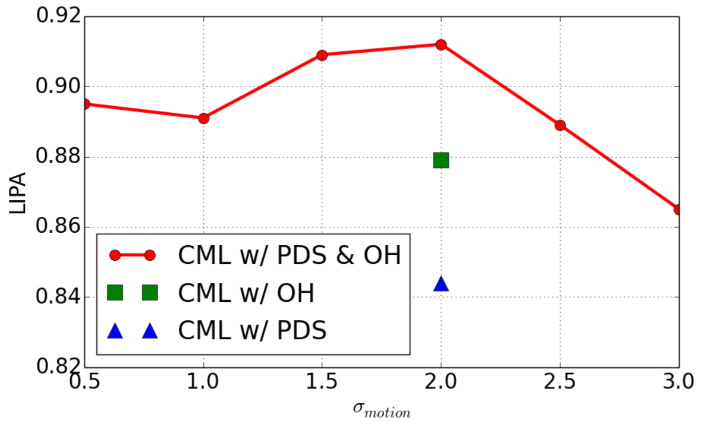


Figure 4.5: Red line with circle markers show the effect of varying motion sigma σ_{motion} in combined color and motion likelihood `CML`. Green square and blue triangle markers show the effect of omitting Player Detection Score (`PDS`) multiplication and Occlusion Handling (`OH`) steps in the tracking methodology. Local Identity Preserving Accuracy (LIPA) is shown on *FB-GS-Tracking Dataset*.

Section 3.4.4.2) is applied during particle distribution; and all likelihood values are combined with Player Detection Score (`PDS`) during particle weighting.

The performance is relatively poor if only color likelihood is used to distribute particles or only motion likelihood used to weight the particles. Color and motion features complement each other in the likelihood model to track more targets through occlusions. As the global likelihood evaluation model, color × motion gives the best results. However, using only color features (with `PDS`) for weighting particles, performs better than color × motion features. Employing motion scores in weighting the associated particles of a track, increase the probability of the tracker to fail in handling non-linear motion patterns. Hence, appearance is a stronger cue to be utilized as the particle weighting model. In the following, using color × motion scores for global likelihood evaluation and color scores for particle weighting will be referred as Color and Motion Likelihood (`CML`).

Figure 4.5 shows that `CML` performs the best when color and motion features are combined with a motion sigma value $\sigma_{motion} = 2$. Increasing the motion sigma further, flattens the Gaussian-shaped likelihood model and decreases the effect of motion in `CML`, and hence drops the accuracy (recall the results of using only the

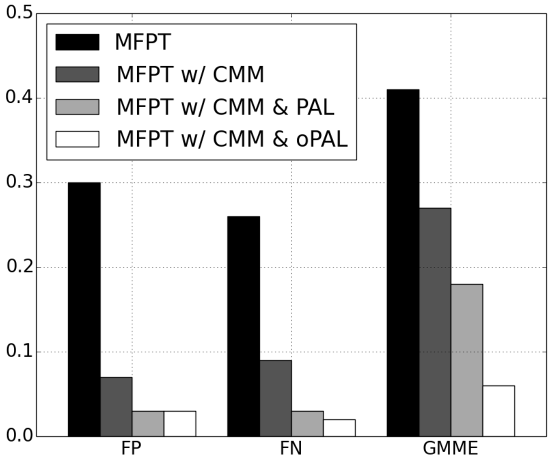


Figure 4.6: Changes in False Positive rate (FP), False Negative rate (FN) and global miss-match rate (GMME) when player identification methods (`CMM`: Collective Motion Model, `PAL`: Positional Appearance Learning that tags new tracks, `oPAL`: Positional Appearance Learning that overrides existing tags) are applied on top of our Model Field Particle Tracking (`MFPT`) algorithm. Experiments are made on *FB-GS-Tracking Dataset*.

color model in the global likelihood evaluation). Figure 4.5 also emphasizes the importance of occlusion handling, which enables assigning particles to the tracks having the highest motion likelihood. If `OH` step is omitted, occluded tracks lose their particles due to low overall likelihood and LIPA drops from 0.912 to 0.844, resulting in 45 more track losses. While weighting the particles of a track, if `PDS` is omitted and only color likelihood is used, LIPA is decreases by 0.033. Since `PDS` gives higher weights to the particles having players centered in the corresponding bounding box on the image plane, it balances color likelihood and helps in avoiding drifts that cause track losses.

4.6.2 Effect of Player Identification

In this section, we measure the effect of adding player identification steps (explained in Section 3.5) on top of our baseline multi-player tracking methodology `CML w/ PDS & OH`. In the following, `CML w/ PDS & OH` will be shortly referred as

Steps of Algorithm	**LIPA**
CML w/ PDS & OH (MFPT)	0.912
MFPT w/ CMM	0.936
MFPT w/ CMM & PAL	0.968
MFPT w/ CMM & oPAL	0.982

Table 4.4: Changes in Local Identity Preserving Accuracy (LIPA) when player identification methods (CMM: Collective Motion Model, PAL: Positional Appearance Learning that tags new tracks, oPAL: Positional Appearance Learning that overrides existing tags) are applied on top of our Model Field Particle Tracking (MFPT). Experiments are made on *FB-GS-Tracking Dataset*.

Model Field Particle Tracking (MFPT). GMOTA components are used to evaluate the performance of Collective Motion Model (CMM) and Positional Appearance Learning (PAL) on *FB-GS-Tracking Dataset*. Two versions of PAL are used in the experiments, in regular PAL, a track is only tagged if it does not contain an identity tag (generally new tracks); the second version oPAL, is authorized to override existing tags if an incorrect identity tag is detected on a track.

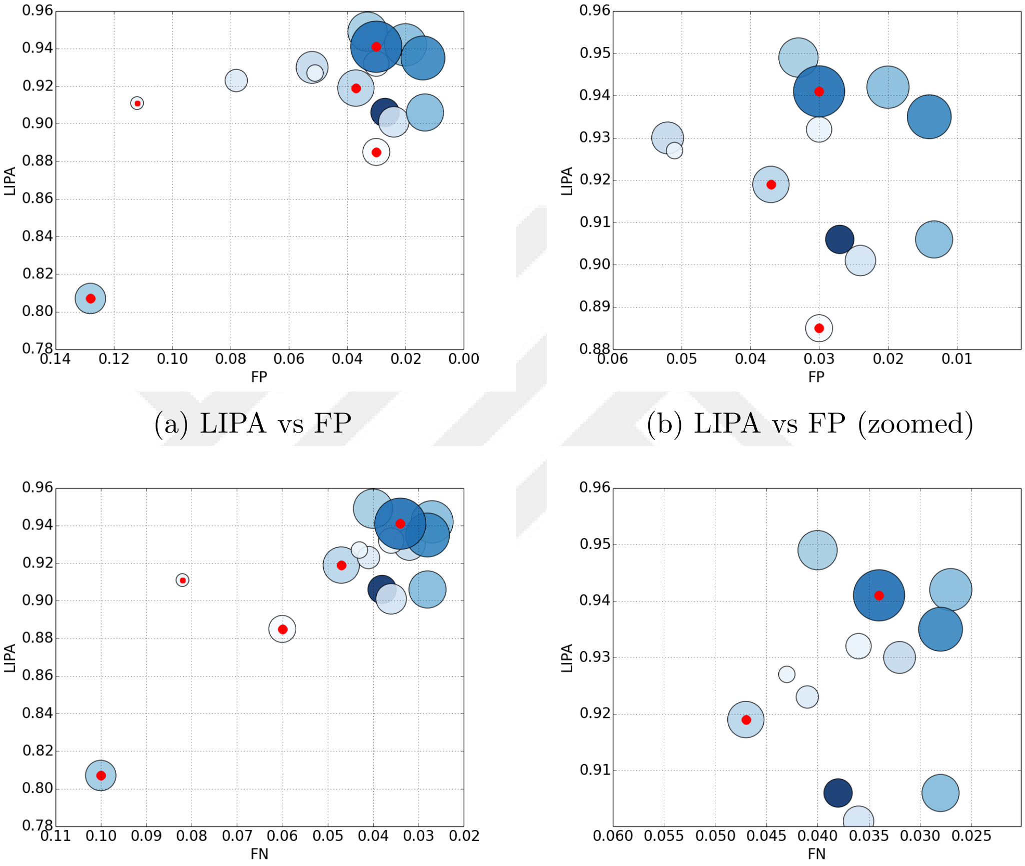
The failure scenarios of the proposed tracking methodology include challenging cases of three or more players being involved in a possession challenge or a tackle in which some players may be completely occluded. In MFPT, a noticeable amount of identity losses occur due to tracks being lost during some of the full occlusions. When the lost player is observed again as a new track after the occlusion, CMM recovers the identity of the track. As seen on Figure 4.6, short-term track identity recovery by MFPT w/ CMM decreases FP, FN and GMME to 0.07, 0.09 and 0.27, respectively. Those tracks that cannot be resolved in the short-term by CMM, are assigned identity tags by PAL, whenever a confident positional appearance classification is made. It follows that MFPT w/ CMM & PAL further reduces the FP, FN and GMME to 0.03, 0.03 and 0.18, respectively. Moreover, if positional appearance learning is authorized to detect wrong identities and correct the tags on the tracks, the global miss-match is significantly reduced to a GMME rate of 0.06. Also, observe in Table 4.4 that how LIPA increases when each player identification step is added on the of the baseline tracking methodology.

4.6.3 Evaluation on Larger Scale

FP, FN and LIPA metrics are used to evaluate the proposed algorithm on a larger scale using the *Sparsely-Annotated-Tracking Dataset*. Performance is measured for 15 professional soccer matches, played on different times of the day in different stadiums, in which the cameras are placed on heights between 9.4 and 28.3 meters above the ground and at distances between 24.4 and 59.1 meters away from the soccer field.

The scatter plots in Figure 4.7 analyze changes in LIPA vs FP and LIPA vs FN with respect to the camera location and game time. In majority of the matches, LIPA exceeds 0.9, FP and FN rates are below 0.05 indicating an accurate tracking performance. Figure 4.7 shows that camera location is influential on the tracking performance. LIPA increases while FP and FN decrease as the camera height above the ground increases. Moreover, placing cameras too close to the field or too far away from the field, negatively affect the tracking performance. When cameras are placed to capture the soccer field from 22.5 meters above the ground and 39.4 meters away from the field, occlusions are better handled from the viewpoint and the highest LIPA of **0.949** is achieved with a low FP and FN rate.

Roof shadow falling on the field in those matches that are played under daylight, causes images to be captured with underexposed and overexposed regions having strong contrast differences. Player detection is negatively affected at some regions of the field. Pixels belonging to the players may not be classified as foreground in the underexposed regions especially if the team is wearing a dark jersey. In such cases, particles occupied by the players may not even be marked as a candidate particle ($\mathbf{s}^m \in \mathbf{S}'$). It is possible to cope with these situations by applying f-HOG classification to all the particles with a severe computational cost. Moreover some tracks may be lost if occlusions occur while passing between regions with contrast differences, since the appearance model cannot adopt immediately to the changes, and a tracker may drift towards an occluding player who is in the region being leaved. As shown in Figure 4.7c, the miss-rate (FN) increases in matches played under daylight. However, LIPA is not negatively affected as



(a) LIPA vs FP

(b) LIPA vs FP (zoomed)

(c) LIPA vs FN

(d) LIPA vs FN (zoomed)

Figure 4.7: Evaluation of the proposed methodology by Local Identity Preserving Accuracy (LIPA), False Positive (FP), and False Negative (FN) metrics on *Sparsely-Annotated-Tracking Dataset* consisting of 15 different soccer matches played in different environmental conditions. Each circle represents a unique 90-minute match. The area of a circle is directly proportional to the height of a camera ranging between 9.4 and 28.3 meters. Circles grow as the camera height increases. The color of a circle gets darker as the distance between the camera and the nearest point of the soccer field increases. Distances vary between 24.4 to 59.1 meters. Circles having a red circle in the middle are matches played in daylight (with roof shadows), while others are night matches. The plots in (c) and (d) are zoomed version of (a) and (b) for better visualization of the distinctions.

Particle Count	FP	FN	GMME	MOTP	T-exec	PP-exec
1 per m^2 $\approx$ 7140	0.10	0.11	0.23	0.53m	9ms	40ms
2x2 per m^2 $\approx$ 28560	0.03	0.02	0.06	0.51m	20ms	41ms
3x3 per m^2 $\approx$ 64260	0.02	0.02	0.05	0.49m	39ms	43ms

Table 4.5: Comparison of False Positive (FP), False Negative (FN), global missmatch rate (GMME), Multiple Object Tracking Precision (MOTP) [65] , and execution times with respect to the particle configuration. T-exec: Execution time per frame of the player detection and tracking algorithm. PP-exec: Execution time per frame of the preprocessing stages of the algorithm. MOTP is the total error (meters) in estimated position for matched object-hypothesis pairs over all frames, averaged by the total number of matches made.

much as FN, because the lost tracks are recovered immediately by the player identification methods.

4.6.4 Computational Cost

The algorithm is implemented in C++ with best-effort optimizations, multithreaded image processing and GPU usage. The two-camera system runs on a laptop with an Intel i7-4710HQ CPU with four cores and eight threads at 2.50GHz. The image acquisition, exposure and light adjustment, foreground extraction and color histogram bin index calculation for each camera are referred as preprocessing steps. These steps execute on different threads and have a total run time of 41 ms per frame. The player classification, player identification and the tracking algorithm, which merge the data from the separate camera threads, execute in parallel with a total throughput time of 20 ms. The execution times per frame are listed in Table 4.5 for different particle configurations. The results in the experiments are reported using an optimal configuration in which 2×2 particles are sampled from each m^2. When 3×3 particles are sampled from each m^2, FP, FN and GMME rates decrease negligibly and the tracking precision increases by only 2 cm, while the tracking execution time increases by a factor of two. The tracking results are poor when only a single particle is sampled from each m^2.

Tracking Algorithm	**LIPA**	**FPS**
MFPT (Our Method)	0.912	16.4
Struct [37]	0.889	4.3
MIL [67]	0.818	2.2
PF w/ CML	0.785	7.2
OAB [68]	0.782	4.6
MPF [49]	0.758	8
CPF [47]	0.723	12.8
CSK [69]	0.665	16.5
Median-Flow [70]	0.609	5.3
KMS [35]	0.595	22.2

Table 4.6: Comparison of our tracking algorithm with the state-of-the-art single-object trackers on the *FB-GS-Tracking Dataset* using Local Identity Preserving Accuracy (LIPA) metric and algorithm execution time (FPS: frame per second).

4.7 Comparison to Related Methods

4.7.1 Comparison with Single-Object Trackers

We compare the proposed multi-player tracking methodology (MFPT), without the player identification steps, to the state-of-the-art single-object trackers on the *FB-GS-Tracking Dataset*. LIPA metric is used to measure the ability of the methods in successfully tracking targets through occlusions by preserving their identities. Tracks are in low-resolution, have similar appearances and move with fast non-linear motion. OpenCV implementations of CSK [69], KMS [35], Median-Flow [70], MIL [67] and OAB [68]; and publicly available source code of Struct [37] were used in the experiments. Color Particle Filter (CPF) and Mixture Particle Filter (MPF) were implemented as described in [47] and [49]. A separate tracker is initiated for each player in every 20 seconds and the number of identity loses are measured at the end of each 20-second interval. LIPAs are listed for each tracker on Table 4.6.

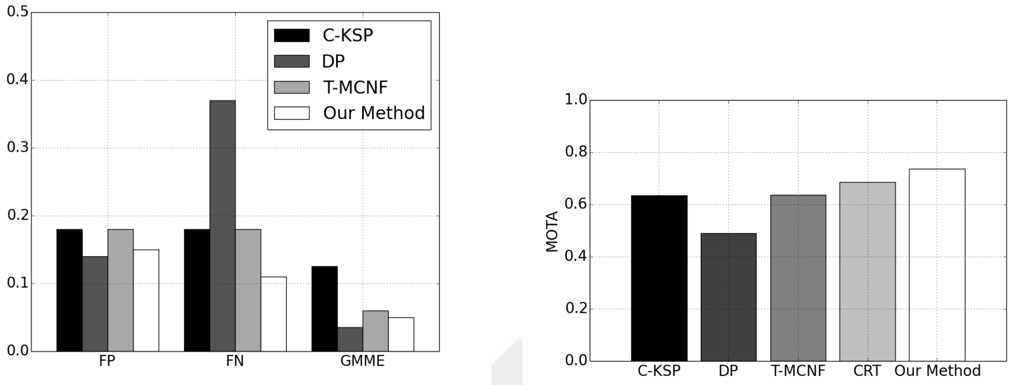
Instead of employing model field particles, using our proposed likelihood function CML within a particle filtering framework (PF w/ CML) on the image plane, achieves a higher accuracy than CPF and MPF. Combining color and motion cues in particle weighting (PF w/ CML), avoids drifting towards other tracks with similar colors, and results in 41 less identity hijackings compared to the standard particle filtering. More importantly, the results demonstrate that the major contribution is the concept of model field particles. Rather than using CML within a particle filtering framework, utilizing model field particles for tracking with the same likelihood function (MFPT) gives a higher LIPA by a rate of 0.127. The proposed framework grasps the global state of the game and interactions among the players by assigning particles to the tracks with respect to global likelihood; handles occlusions on the ground plane by allowing particles to be shared by two or more tracks; handles non-linear motion patterns by combining appearance features with player detection scores at particle weighting step; and achieves 84 less track losses than PF w/ CML.

LIPA is lower than 0.665 for CSK, Median-Flow and KMS, since these trackers are more suitable for very smooth and predictable movements when the objects are visible throughout the whole sequence. In these trackers, majority of the losses are due to fast and non-linear motion of the players. The OAB and MIL algorithms, which use online classifiers to separate the target from the background and surroundings, show similar performances with an accurate tracking on the dataset. However, when tracking multiple players having similar appearances in low-resolution and when the background is homogeneous such as the grass, the online classifiers in these methods learn generic discriminative functions that can distinguish players from the grass, but may fail to distinguish teammates when they are near each other or in an occlusion. So, despite the heavy computational load of OAB and MIL, the results are still lower than MFPT by 0.13 and 0.094. Struct tracking algorithm uses a kernelized structured output SVM, which is learned online to provide adaptive tracking. Struct (one of the best performing algorithms in a benchmarking study [71]) shows the closest tracking performance to our method and is only lower by 0.023 with respect to LIPA. However, our approach may be favorable to Struct in terms of efficiency reflected by the execution time. In summary, results show that when compared to the state-of-the-art single-object trackers using different approaches, the proposed likelihood model and the concept of model field particles show solid tracking performance to handle low-resolution targets, fast non-linear motion, and occlusions.

4.7.2 Comparison with Multi-Player Tracking Methods

We compare our complete methodology `MFPT w/ CMM & PAL` with the state-of-the-art tracking methods that reported results on the benchmark *ISSIA-Tracking Dataset*. Fig. 4.8a compares the provided tracking errors (in [42]) of K-Shortest Path (C-KSP) tracker with appearance [72], Dynamic Program (DP) tracking algorithm similar to the ones in [73, 74], and Tracklet-based Multi-Commodity Network Flow (T-MCNF) [42] with `MFPT w/ CMM & PAL`. In Figure 4.8b, our method is compared using MOTA with C-KSP, DP, T-MCNF, and Contextual Reasoning Tracking (CRT) [41]. Although a combined MOTA for the whole system is not reported in [41], the results for the individual cameras are averaged for CRT.

We only mapped three cameras in a single-view configuration to our model field and generated field particles, as explained in Sec. 3.2.4. Note that there is only a little overlap between the cameras and some portion of the field is not visible in the single-view configuration. This introduces additional errors when tracks are moving between the cameras. C-KSP, DP and T-MCNF utilize the multi-view configuration of the *ISSIA-Tracking Dataset*. Although the formulation is a lot more complex, utilizing a multi-view configuration to detect the players in each frame, representing sequence of detections as graph nodes and linking nodes to extract player trajectories by examining a batch of frames are very strong cues for accurately resolving occlusions and preserving identities of the tracks. Trajectories of the players may be more meaningful when they are examined in space-time, rather than trying to predict the next position of the players by only looking at the current situation of the game. In contrast to processing batch of frames, in model field particles tracking, game state transitions are efficiently described as a first-order Markov process and players are tracked using the available information only in the current and the previous frame without introducing any time delays.



(a) (b)

Figure 4.8: Comparison of the proposed multi-player tracking and player identification approach `MFPT w/ CMM & PAL` (Our Method) with the other multi-object tracking algorithms that reported results over the publicly available *ISSIA-Tracking Dataset* [64]. C-KSP: K-Shortest Path tracker with appearance [72]. DP: Tracking with Dynamic Programming algorithm similar to the ones in [73, 74]. T-MCNF: Tracklet-based Multi-Commodity Network Flow [42]. CRT: Contextual Reasoning Tracker [41]. (a) False Positive (FP), False Negative (FN), and global miss-match (GMME) rates are compared; (b) Multiple Object Tracking Accuracy (MOTA) is compared for different methods.

Despite the fundamental differences in the approaches and the disadvantages of the single-view configuration, our results are comparable to those of the state-of-the-art multi-view tracking methods which process frames as a batch and require larger batch sizes for better performance in global trajectory optimization on POMs [73]. The proposed approach achieves a higher MOTA of **0.737** compared to the other multi-player tracking methods. Moreover, majority of the FP and FN are due to the regions that are not visible from the single-view. However, the FP and FN rates are still lower in our approach. Finally, as reflected by the GMME rate, our approach shows a similar identity-preserving tracking performance to the multi-view, data-association and global optimization-based methods.

Chapter 5

Conclusions

5.1 Summary

We introduce the concept of model field particles that is specifically designed to track interacting players with similar appearances. Particles are sampled at fixed positions, and each particle on the model soccer field is represented by a bounding box on the image plane. Tracks are detected on the particles by foreground extraction and a supervised player classifier that is specifically trained for soccer. Players are tracked through challenging occlusions by globally evaluating the likelihood of the tracks being on the model field particles. A combined color and motion model is used to calculate track-to-particle likelihoods and distribute the particles among the tracks. To precisely estimate the position of the tracks, the associated particles are weighted by a color model that is supported with a player detection score. Positional appearance of the players is learned online by SVM classifiers using relative position and raw-pixel features. Positional appearance learning detects incorrect identities and modifies the tags on the tracks. Moreover, track losses are recovered by assigning the correct identities to the new observations by the collective motion model in the short-term and by the positional appearance learning in the long-term.

Experimental results show that color and motion features are complementary in tracking multiple players having similar appearances. Utilizing our combined color and motion model within the concept of model field particles, significantly improves the tracking performance compared to using the same model as a function of the standard Particle Filtering framework. The proposed framework gives better results in preserving identities during tracking when compared to the state-of-the-art single-object trackers on our dataset.

Adding player identification steps on top of the tracking algorithm further increase tracking accuracy. Collective motion model reduces false positive and false negative rates; whereas, the positional appearance learning decreases global identity miss-match rate. The complete approach was also evaluated on larger scale using sparsely annotated matches played in different environmental conditions. Accurate tracking results with low number of identity switches, false positives and false negatives were obtained in majority of the matches. Camera height was found to be influential on the tracking results, and the miss rate was higher in matches played in daylight with the stadium roof shadow falling on the field. Finally, the effectiveness of the approach was demonstrated on a publicly available dataset by comparing our tracking results to the other multi-player tracking methods.

5.2 Discussion

Tracking multiple players in sports is as challenging as any other multi-object tracking problem. Targets are in low-resolution, they look almost identical, they frequently occlude each other and they move in non-linear motion patterns. The state-of-the-art trackers use online classifiers to separate targets from the background and track them using discriminative models. In sports, low-resolution players with similar appearances and the homogeneous background cause generic discriminative functions to be learned for each target. These models fail to distinguish tracks when they are near each other or involved in an occlusion, resulting in track losses and identity hijackings. Hence, combining simpler features such as color and motion, as in our approach, is more suitable for multi-player tracking in the sports domain.

Rather than relaying on a strong discriminative model, multi-player tracking can be formulated as a data-association problem in which players are detected in each frame and detections are linked to form the target trajectories. Frames are processed as a bulk trying to reach a global optimal solution by moving forward and backward in time. Rather than trying to predict the next position of the players by only looking at the current state, examining trajectories in space-time is more meaningful in resolving complex motion patterns and occlusions. However, such approaches introduce complicated formulations and require larger batch sizes for better performance in global trajectory optimization.

In order to simply the formulation, Markovian property is exploited to employ separate probabilistic trackers for each target. The particle filtering approaches generate many particles on the image plane around the target for tracking. However, tracking multiple targets with similar appearances cause particle degeneration, in which particles of a track are propagated towards another target or transferred to another mixture component. Hence, identity switches or hijackings occur among tracks during occlusions and the performance is not as solid as the data-association methods.

As in our approach, if particles are densely sampled at fixed positions on the real-world ground plane, few particles are needed to accurately track the target resulting in a more efficient tracker and tracking processes is simplified such that there is no need for a particle re-sampling step compared to the standard particle filtering framework. Moreover, occlusions are handled implicitly, since an occlusion may only occur on the image plane and tracks cannot be on top of each other on the ground plane. The experimental results reflect that the concept of model field particles is more suitable for tracking multiple targets with similar appearances on a calibrated ground than the standard particle filtering approaches and object trackers that employ different features in online discriminative learning. In addition to its efficiency and simpler formulation, the performance of our approach in sports player tracking is as good as the multi-view, data-association and global optimization-based methods.

5.3 Future Work

In the near future, we plan to replace full-HD cameras with 4K cameras (3840×2160) to significantly improve the image quality. Increased image resolution will result in each player to be represented with more number of pixels and hence improve the accuracy of multi-player tracking and appearance learning.

As a feature direction, the video and tracking data that was collected for a whole soccer league season can be utilized in advanced learning models. Deep learning networks can be used for building a more accurate player detector. Moreover, the deep learning based player detection network can also be used as a feature extractor for training an online player identification model. Similarly, seasonal tracking data including xy-positions of all the players can be used to extract parameters for an autoregressive motion model or train a network to predict the next position of tracks given the priors.

The proposed algorithm was already deployed in proof of concept applications to track players in other sports such as basketball, rugby, and field hockey. We plan to build upon the encouraging initial results, prove the effectiveness of the algorithm and generalize it to work in other sports as well. Finally, the proposed approach can be extended to a multi-view version for those sports that cannot be handled by a single-view camera configuration and additional pan-tilt-zoom cameras can be installed to read jersey numbers for automated player identification.

Appendix A

Demo of the Tracking

We provide demonstration material for the proposed multi-player tracking algorithm.

i. Video in the following link illustrates likelihood of a player being on the densely sampled particles with respect to different models (Video version of Figure 3.7): `https://youtu.be/xCgE0wSG9XI`.

ii. Video in the following link illustrates distributing particles among the tracks and weighting particles for updating track locations (Video version of Figure 3.8): `https://youtu.be/f1HTEZBrhoA`.

iii. Video in the following link shows the output of the tracking on *FB-GS-Tracking Dataset*: `https://youtu.be/a3LjiSCSO-k`.

iv. Video in the following link shows the output of the tracking on *ISSIA-Tracking Dataset*: `https://youtu.be/-kLEdIvjh1U`.

v. Video in the following link shows output of the tracking on different soccer matches from the Turkish Super League played in 2015-2016 season: `https://youtu.be/M51C9nViDO4`.

vi. Figure A.1 shows players being tracked in a sequence of images by `MFPT w/ CMM & PAL`.

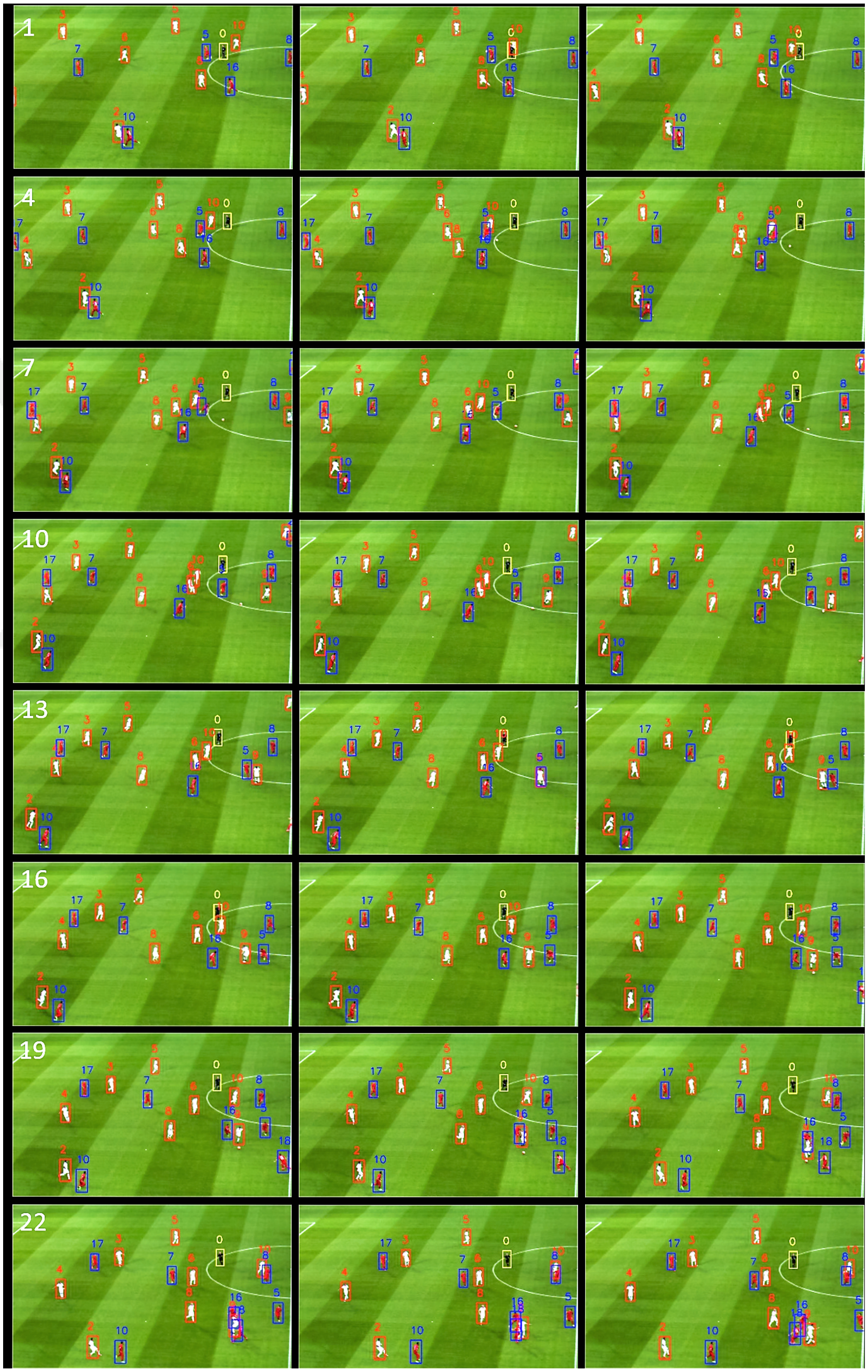


Figure A.1: Players being tracked in a sequence of images. Frame number is given on the top left. Illustration is in 5 FPS.

Bibliography

[1] T. D'Orazio and M. Leo, "A review of vision-based systems for soccer video analysis," *Pattern Recognition*, vol. 43, no. 8, pp. 2911–2926, 2010.

[2] Fédération Internationale de Football Association (FIFA), "Big count." http://www.fifa.com/worldfootball/bigcount/index.html, 2006.

[3] Sentio Sports Analytics, "Sentio match report." http://www.sentiosports.com/, 2016.

[4] S. Baysal and P. Duygulu, "Sentioscope: A soccer player tracking system using model field particles," *IEEE Transactions on Circuits and Systems for Video Technology*, 2015.

[5] A. W. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, "Visual tracking: An experimental survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 36, no. 7, pp. 1442–1468, 2014.

[6] X. Li, W. Hu, C. Shen, Z. Zhang, A. Dick, and A. V. D. Hengel, "A survey of appearance models in visual object tracking," *ACM Transactions on Intelligent Systems and Technology*, vol. 4, no. 4, 2013.

[7] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Computing Surveys*, vol. 38, no. 4, p. 13, 2006.

[8] P. F. Gabriel, J. G. Verly, J. H. Piater, and A. Genon, "The state of the art in multiple object tracking under occlusion in video sequences," in *Advanced Concepts for Intelligent Vision Systems*, pp. 166–173, Citeseer, 2003.

[9] D. M. Gavrila, “The visual analysis of human movement: A survey,” *Computer Vision and Image Understanding*, vol. 73, no. 1, pp. 82–98, 1999.

[10] J. K. Aggarwal and Q. Cai, “Human motion analysis: A review,” in *Nonrigid and Articulated Motion Workshop*, pp. 90–102, IEEE, 1997.

[11] A. Dearden, Y. Demiris, and O. Grau, “Tracking football player movement from a single moving camera using particle filters,” in *Proceedings of European Conference on Visual Media Production (CVMP)*, pp. 29–37, 2006.

[12] S. Gedikli, J. Bandouch, N. von Hoyningen-Huene, B. Kirchlechner, and M. Beetz, “An adaptive vision system for tracking soccer players from variable camera settings,” in *Proceedings of International Conference on Computer Vision Systems (ICVS)*, 2007.

[13] M. Herrmann, M. Hoernig, and B. Radig, “Online multi-player tracking in monocular soccer videos,” in *Proceedings of AASRI Conference on Sports Engineering and Computer Science*, pp. 30–37, Elsevier, 2014.

[14] M.-C. Hu, M.-H. Chang, J.-L. Wu, and L. Chi, “Robust camera calibration and player tracking in broadcast basketball video,” *IEEE Transactions on Multimedia*, vol. 13, no. 2, pp. 266–279, 2011.

[15] J. Liu, X. Tong, W. Li, T. Wang, Y. Zhang, and H. Wang, “Automatic player detection, labeling and tracking in broadcast soccer video,” *Pattern Recognition Letters*, vol. 30, no. 2, pp. 103–113, 2009.

[16] W.-L. Lu, J.-A. Ting, J. J. Little, and K. P. Murphy, “Learning to track and identify players from broadcast sports videos,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 35, no. 7, pp. 1704–1716, 2013.

[17] H. Ok, Y. Seo, and K. Hong, “Multiple soccer players tracking by condensation with occlusion alarm probability,” in *International Workshop on Statistically Motivated Vision Processing*, 2002.

[18] 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,” *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1652–1667, 2011.

[19] P. Figueroa, N. Leite, R. M. Barros, I. Cohen, and G. Medioni, “Tracking soccer players using the graph representation,” in *Proceedings of International Conference on Pattern Recognition (ICPR)*, pp. 787–790, IEEE, 2004.

[20] M. Kristan, J. Perš, M. Perše, and S. Kovačič, “Closed-world tracking of multiple interacting targets for indoor-sports applications,” *Computer Vision and Image Understanding*, vol. 113, no. 5, pp. 598–611, 2009.

[21] T. Misu, M. Naemura, W. Zheng, Y. Izumi, and K. Fukui, “Robust tracking of soccer players based on data fusion,” in *Proceedings of International Conference on Pattern Recognition (ICPR)*, pp. 556–561, IEEE, 2002.

[22] C. J. Needham and R. D. Boyle, “Tracking multiple sports players through occlusion, congestion and scale,” in *Proceedings of British Machine Vision Conference (BMVC)*, pp. 93–102, 2001.

[23] M. Schlipsing, J. Salmen, M. Tschentscher, and C. Igel, “Adaptive pattern recognition in real-time video-based soccer analysis,” *Journal of Real-Time Image Processing*, pp. 1–17, 2014.

[24] H. Ben Shitrit, J. Berclaz, F. Fleuret, and P. Fua, “Tracking multiple people under global appearance constraints,” in *Proceedings of International Conference on Computer Vision (ICCV)*, pp. 137–144, IEEE, 2011.

[25] S. Iwase and H. Saito, “Parallel tracking of all soccer players by integrating detected positions in multiple view images,” in *Proceedings of International Conference on Pattern Recognition (ICPR)*, pp. 751–754, IEEE, 2004.

[26] M. Leo, N. Mosca, P. Spagnolo, P. L. Mazzeo, T. D’Orazio, and A. Distante, “Real-time multiview analysis of soccer matches for understanding interactions between ball and players,” in *Proceedings of the International Conference on Content-based Image and Video Retrieval*, pp. 525–534, ACM, 2008.

[27] R. Martín and J. M. Martínez, "A semi-supervised system for players detection and tracking in multi-camera soccer videos," *Multimedia Tools and Applications*, vol. 73, no. 3, pp. 1617–1642, 2014.

[28] E. Morais, A. Ferreira, S. A. Cunha, R. M. Barros, A. Rocha, and S. Goldenstein, "A multiple camera methodology for automatic localization and tracking of futsal players," *Pattern Recognition Letters*, vol. 39, pp. 21–30, 2014.

[29] E. Morais, S. Goldenstein, A. Ferreira, and A. Rocha, "Automatic tracking of indoor soccer players using videos from multiple cameras," in *Proceedings of Conference on Graphics, Patterns and Images (SIBGRAPI)*, pp. 174–181, IEEE, 2012.

[30] M. Xu, J. Orwell, and G. Jones, "Tracking football players with multiple cameras," in *International Conference on Image Processing*, vol. 5, pp. 2909–2912, IEEE, 2004.

[31] R. T. Collins and P. Carr, "Hybrid stochastic/deterministic optimization for tracking sports players and pedestrians," in *European Conference on Computer Vision (ECCV)*, pp. 298–313, Springer, 2014.

[32] J. Liu and P. Carr, "Detecting and tracking sports players with random forests and context-conditioned motion models," in *Computer Vision in Sports*, pp. 113–132, Springer, 2014.

[33] J. Liu, P. Carr, R. T. Collins, and Y. Liu, "Tracking sports players with context-conditioned motion models," in *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1830–1837, IEEE, 2013.

[34] Y. Seo, S. Choi, H. Kim, and K.-S. Hong, "Where are the ball and players? soccer game analysis with color-based tracking and image mosaick," in *Image Analysis and Processing*, pp. 196–203, Springer, 1997.

[35] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 24, no. 5, pp. 603–619, 2002.

[36] J. Perš and S. Kovačič, "Tracking people in sport: Making use of partially controlled environment," in *Computer Analysis of Images and Patterns*, pp. 374–382, Springer, 2001.

[37] S. Hare, A. Saffari, and P. H. Torr, "Struck: Structured output tracking with kernels," in *Proceedings of International Conference on Computer Vision (ICCV)*, pp. 263–270, IEEE, 2011.

[38] A. Li, F. Tang, Y. Guo, and H. Tao, "Discriminative nonorthogonal binary subspace tracking," in *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 258–271, Springer, 2010.

[39] R. L. Streit and T. E. Luginbuhl, "Maximum likelihood method for probabilistic multihypothesis tracking," in *Symposium on Optical Engineering and Photonics in Aerospace Sensing*, pp. 394–405, 1994.

[40] Y. Bar-Shalom and T. Fortman, *Tracking and Data Association.* 1988.

[41] R. Di Lascio, P. Foggia, G. Percannella, A. Saggese, and M. Vento, "A real time algorithm for people tracking using contextual reasoning," *Computer Vision and Image Understanding*, vol. 117, no. 8, pp. 892–908, 2013.

[42] H. Ben Shitrit, J. Berclaz, F. Fleuret, and P. Fua, "Multi-commodity network flow for tracking multiple people," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 36, no. 8, pp. 1614–1627, 2014.

[43] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174–188, 2002.

[44] M. Isard and J. MacCormick, "Bramble: A Bayesian multiple-blob tracker," in *Proceedings of International Conference on Computer Vision (ICCV)*, pp. 34–41, IEEE, 2001.

[45] J. Czyz, B. Ristic, and B. Macq, "A color-based particle filter for joint detection and tracking of multiple objects," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 217–220, IEEE, 2005.

[46] R. Hess and A. Fern, "Discriminatively trained particle filters for complex multi-object tracking," in *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 240–247, IEEE, 2009.

[47] P. Pérez, C. Hue, J. Vermaak, and M. Gangnet, "Color-based probabilistic tracking," in *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 661–675, Springer, 2002.

[48] J. Vermaak, A. Doucet, and P. Pérez, "Maintaining multimodality through mixture tracking," in *Proceedings of International Conference on Computer Vision (ICCV)*, pp. 1110–1116, IEEE, 2003.

[49] K. Okuma, A. Taleghani, N. De Freitas, J. J. Little, and D. G. Lowe, "A boosted particle filter: Multitarget detection and tracking," in *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 28–39, Springer, 2004.

[50] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. I–511, 2001.

[51] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 886–893, IEEE, 2005.

[52] N. Nourani-Vatani and J. Roberts, "Automatic camera exposure control," in *Australasian Conference on Robotics and Automation*, pp. 1–6, 2007.

[53] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision.* Cambridge University Press, 2003.

[54] A. Agarwal, C. Jawahar, and P. Narayanan, "A survey of planar homography estimation techniques," tech. rep., IIIT/TR/2005/12 International Institute of Information Technology, 2005.

[55] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 32, no. 9, pp. 1627–1645, 2010.

[56] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in *Proceedings of International Conference on Pattern Recognition (ICPR)*, pp. 28–31, IEEE, 2004.

[57] Z. Zivkovic and F. van der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern Recognition Letters*, vol. 27, no. 7, pp. 773–780, 2006.

[58] E. Osuna, R. Freund, and F. Girosit, "Training support vector machines: an application to face detection," in *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 130–136, IEEE, 1997.

[59] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in Large Margin Classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[60] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *Proceedings of International Conference on Pattern Recognition (ICPR)*, pp. 850–855, IEEE, 2006.

[61] T. J. Broida and R. Chellappa, "Estimation of object motion parameters from noisy images," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 8, no. 1, pp. 90–99, 1986.

[62] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[63] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 24, no. 4, pp. 509–522, 2002.

[64] T. D'Orazio, M. Leo, N. Mosca, P. Spagnolo, and P. L. Mazzeo, "A semi-automatic system for ground truth generation of soccer video sequences,"

in *Proceedings of International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 559–564, IEEE, 2009.

[65] B. Keni and S. Rainer, "Evaluating multiple object tracking performance: the CLEAR MOT metrics," *Journal on Image and Video Processing*, 2008.

[66] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang, "Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 31, no. 2, pp. 319–336, 2009.

[67] B. Babenko, M.-H. Yang, and S. Belongie, "Visual tracking with online multiple instance learning," in *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 983–990, IEEE, 2009.

[68] H. Grabner, M. Grabner, and H. Bischof, "Real-time tracking via on-line boosting," in *Proceedings of British Machine Vision Conference (BMVC)*, p. 6, 2006.

[69] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 702–715, Springer, 2012.

[70] Z. Kalal, K. Mikolajczyk, and J. Matas, "Forward-backward error: Automatic detection of tracking failures," in *Proceedings of International Conference on Pattern Recognition (ICPR)*, pp. 2756–2759, IEEE, 2010.

[71] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2411–2418, IEEE, 2013.

[72] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua, "Multiple object tracking using k-shortest paths optimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 33, no. 9, pp. 1806–1819, 2011.

[73] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, “Multicamera people tracking with a probabilistic occupancy map,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 30, no. 2, pp. 267–282, 2008.

[74] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, “Globally-optimal greedy algorithms for tracking a variable number of objects,” in *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1201–1208, IEEE, 2011.