Computer Vision and Image Understanding 159 (2017) 19–46

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

# A survey on player tracking in soccer videos

M. Manafifard[a],[*], H. Ebadi[b], H. Abrishami Moghaddam[c]

[a] Dept. of Photogrammetry and Remote Sensing, K.N. Toosi University of Technology, Valieasr Street, Tehran, Iran
[b] Dept. of Photogrammetry and Remote Sensing, K. N. Toosi University of Technology, Valieasr Street, Tehran, Iran
[c] Dept. of Electrical Engineering, K. N. Toosi University of Technology, Seyedkhandan Street, Tehran, Iran

ARTICLE INFO

*Article history:*
Received 14 March 2016
Revised 31 January 2017
Accepted 3 February 2017
Available online 4 February 2017

*Keywords:*
Soccer
Playfield
Player
Detection
Labeling
Appearance
Tracking
Occlusion
Evaluation

ABSTRACT

There is a growth of demand for automatically analyzing soccer matches and tactics. Since players are the focus of attentions in soccer matches, player tracking is a fundamental element in most soccer video analysis. The aim of player tracking is to extract the trajectories of players, and its input is provided through some preprocessing steps including playfield detection, player detection, player labeling, occlusion handling and player appearance modeling. Soccer player tracking is a complex and challenging task due to difficulties such as blur, illumination change and heavy occlusions. This paper presents the state-of-the-art in preprocessing and processing methods for soccer player tracking. We categorize different approaches, analyze their strengths and weaknesses, review evaluation criteria and conclude future research directions.

## 1. Introduction

Automatic soccer video analysis is a mandatory response to the growing demand by sport professionals and fans for extracting semantic information. Soccer video analysis has also attracted wide applications, such as player trajectory extraction, content retrieval and indexing, summarization, highlight detection, 3D reconstruction of the soccer match, animations, virtual view generation, virtual content insertion, visualization, editorial content creation and content enhancement, content-based video compression, tactical analysis, pattern of attack or goal analysis, statistical evaluations, player action recognition, verification of referee decisions, adapting the training plan and evaluating strengths or weaknesses of a team or a player. Player detection and tracking are fundamental elements required for extracting such understanding of the game. In general, player detection and tracking are quite challenging due to many difficulties, such as similar appearance of players, complex interactions and severe occlusions, unconstrained outdoor environment, changing background, varying number of players with unpredictable movements, abrupt camera motion and zoom, calibration inaccuracy due to the low textured field and edited broadcast video, noise, lack of pixel resolution especially on small distant players, clutter and motion blur. Examples of blurred players and lines are shown in Fig. 1.

An interesting review of the state-of-the-art in tracking algorithms can be found in Yilmaz et al. (2006), and a survey on visual tracking was provided by Yang et al. (2011). Two surveys on soccer video analysis were also provided by D' Orazio and Leo (2010) and Oskouie et al. (2014). However, they were focused on various soccer video applications, while methodologies for player detection and tracking were concisely reviewed. A review of spatio-temporal analysis of team sports was also presented by Gudmundsson and Horton (2016); however, it rarely focused on soccer player tracking. The main aim of this paper is to review in detail player tracking and its preprocessing steps. Moreover, different criteria for evaluation of the performance are reviewed. Accordingly, evaluation, player tracking and its preprocessing steps (playfield detection, player detection, occlusion resolution and appearance modeling) are shown in Fig. 2. Playfield detection plays a primary role in soccer video analysis. It eliminates the spectator region and reduces false alarms and noises within the playfield, which is of much benefit to the subsequent tracking procedure. Moreover, correct player detection is essential for initializing the tracker and provides observations required by some trackers. Another important aspect is player classification into five classes corresponding to two teams, two goalkeepers and referee, namely, player labeling. On the other side, occlusion is the most challenging issue in tracking soccer players, which occurs when

[*] Corresponding author.
*E-mail addresses:* mmanafifard@mail.kntu.ac.ir (M. Manafifard), ebadi@kntu.ac.ir (H. Ebadi), moghaddam@kntu.ac.ir (H. Abrishami Moghaddam).

http://dx.doi.org/10.1016/j.cviu.2017.02.002
1077-3142/© 2017 Elsevier Inc. All rights reserved.

**Fig. 1.** Scenes with blurred players (in particular in feet) and blurred or duplicate lines.

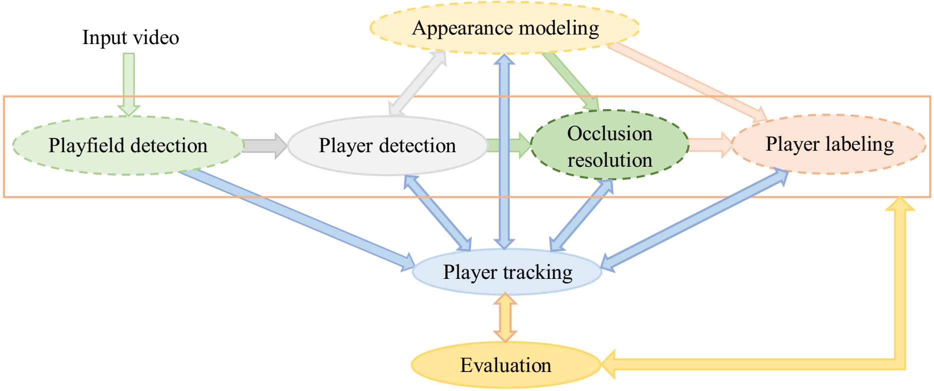


**Fig. 2.** The flowchart of soccer player tracking.

players hide each other either partially or completely. Occlusion is sometimes so severe that even the human observer can hardly see the occluded player, and the low quality of the images complicates more the problem. Also, more difficult situation arises due to the occlusion among teammates having similar appearances. Moreover, most of the player labeling, color-based player detection, occlusion resolution and player tracking methods require appearance representation of the players and referee while appearance modeling faces challenges such as updating models and selecting the most discriminative features. Player detection result can also be used as input to the appearance modeling step (gray bidirectional arrow) by restricting it to the detected player region. At the end of the tracking step, player tracking and its prior steps (rectangle in Fig. 2) are evaluated using some evaluation criteria. However, some prior steps are optional depending on the tracking step (dotted circles). For instance, playfield detection can be ignored, or appearance modeling is not required when color cues are ignored by a tracker. Occlusion resolution can also be performed within the tracking step (e.g. graph tracking), or player labeling can be solved during the tracking based on color cues (e.g. particle filter (PF)). Moreover, tracking results can be used as inputs to some prior steps, such as player detection, player labeling, occlusion resolution and appearance modeling (blue bidirectional arrows). For instance, probable player regions in the current frame can be predicted using the tracking result in the previous frame, or the previous appearance model can be updated using the appearance model of the tracked player. Occlusion situation can also be predicted using the distance between tracked players in the previous frame, or players' labels can be updated regarding the color-based tracking result. Moreover, all these steps can be modified regarding the evaluation results (orange bidirectional arrows).

Motion capture in commercial applications can be achieved with tracking reflective, magnetic markers or global positioning system (GPS) (Rangsee et al., 2013) on a player's body which are not always possible in sport domains, where the player movement can be affected or markers are not allowed. As a solution, computer vision techniques aim to dispose of such markers. Accordingly, some player tracking systems such as TRACAB (Capturing and visualizing large scale human action, 2016) achieved real-time and high precision tracking thanks to advanced camera setups, developed hardware and stereo vision technology.

The player detection and labeling provide observations of players; however, it is necessary to relate observations via a tracker. Moreover, most of the errors resulting from missed detections, false positives or mislabeling can be resolved by incorporating a tracker. A large number of tracking algorithms, such as Kalman filter (KF), PF, meanshift, snake and template matching, have been applied to deal with this topic. Generally, player tracking is performed during two main steps, namely, filtering and data association. Filtering is about unknown state estimation (e.g. position and velocity). However, multiple-player tracking involves the problem of data association for jointly tracking of players, since independent tracking of players tends to fail for closely spaced players. Therefore, data association works out which measurements are generated by which players. The problem gets more challenging with an increase in false alarm rate, missed observations and density of tracks. Accordingly, joint probabilistic data association filter (JPDA) and multiple hypothesis tracking (MHT) are two well-known techniques in the literature.

Although useful information was provided by previous literature surveys, the main drawback was addressing relating works without appropriate categorization of different methods or discussion about their weaknesses and strengths. Accordingly, the survey on different prior steps was superficial in the literature, and it was often addressed partially. Although an informative survey on visual tracking was provided by Yang et al. (2011), few tracking methods were reviewed. In addition, some literature surveys were focused on different soccer (D' Orazio and Leo, 2010; Oskouie et al., 2014) or sport (Gudmundsson and Horton, 2016) video analysis, and methodologies for player tracking and its prior steps were concisely reviewed. In D' Orazio and Leo, (2010) and Oskouie et al. (2014), prior steps, such as playfield detection, player detection, player labeling, appearance modeling, evaluation criteria and different camera setups were not reviewed, or few methods were addressed along with the tracking step. Moreover, different tracking methods for soccer player tracking were not categorized, and few trackers were reviewed. The main goal of this paper is to review different preprocessing steps for soccer player tracking, categorize different tracking frameworks and compare them in terms of the available evaluation criteria. Accordingly, separate sections are dedicated to different prior steps and tracking methods. It may help researchers to get familiar with the renowned and state-of-the-art methods in the domain. Moreover, it highlights future research directions to compensate for the weaknesses and low performances of the available algorithms. This paper also gives insight into enhancing existing trackers or proposing new ones, and each reviewed tracker can be developed in different applications.

This paper is organized as follows. In Section 2, input videos including different camera setups are described. Preprocessing methods (i.e. playfield detection, player detection, player labeling and appearance modeling), which provide inputs to the tracking step, are reviewed in Sections 3–5, respectively. The review of player tracking methods is described in Section 6. The related works on occlusion resolution are explained in Section 7. The evaluation criteria are presented in Section 8, and finally research challenges and future research directions are presented in Section 9.

## 2. Input videos

Soccer videos can be categorized into those captured by static, dynamic/moving and mixed cameras where the extracted data from each setup category was obtained through single or multiple cameras (multiple stationary cameras (Xu et al. 2004a, Choi and Seo, 2011; Ren et al., 2009; Khan and Shah, 2009; D'Orazio et al., 2009; Iwase and Saito, 2004; Martín and Martínez, 2013; Montañés Laborda et al., 2011; Junior and Anido, 2004; Figueroa et al., 2006; Sullivan and Carlsson, 2006; Iwase and Saito, 2003; Inamoto and Saito, 2007; Barros et al., 2007; Taki et al., 1996; Baysal and Duygulu, 2016; Ohno et al., 2000; Kasuya et al., 2008; Iwase and Saito, 2002; Misu et al., 2004; Hamid et al., 2010; Abbott and Williams, 2007; Sullivan et al., 2009; Mentzelopoulos et al., 2012; Joo and Chellappa, 2007; Enomoto and Saito, 2009; Poppe et al., 2010), single stationary camera (Needham and Boyle, 2001; Vandenbroucke et al., 1997a, Vos and Brink, 2009; Zhong et al., 2006; Vandenbroucke et al., 2003), moving camera including broadcast streams (Liu et al., 2009; Beetz et al., 2006; Beetz et al., 2007; Pallavi et al., 2008; Dearden et al., 2006; Lefèvre et al., 2000; Mackowiak, 2013; Xing et al., 2011; Barceló et al., 2005; Chiang et al., 2009; Kim et al., 2003; Maćkowiak et al., 2010; Heydari and Moghadam, 2012; Utsumi et al., 2002; Huang et al., 2007; Naemura et al., 2000; Mochizuki et al., 2009; Khatoonabadi and Rahmati, 2009; Choi et al., 2004; Sato and Aggarwal, 2005; Vermaak et al., 2005; Yoon et al., 2002; Seo et al., 1997; Intille and Bobick, 1995a, Zhang et al., 2008; Zhu et al., 2006), multiple moving cameras (Hayet et al., 2005; Hoyningen-Huene and Beetz, 2009) and mixed cameras (Misu et al., 2009)).

A large number of trackers, such as Markov chain Monte Carlo (MCMC) (Liu et al., 2009), template matching (Yoon et al., 2002), template matching and merge-split (Khatoonabadi and Rahmati, 2009), template matching and closed world method (Intille and Bobick, 1995a), KF and template matching (Seo et al., 1997), mixture Kalman to solve joint probabilistic data association (Vermaak et al., 2005), KF and graph (Barceló et al., 2005), KF and point distribution manifolds (Mathes and Piater, 2006), sequential Monte Carlo joint probabilistic data association (SMCJPDA) (Zheng and Xue, 2009), improved meanshift (Chiang et al., 2009), dual-mode two-way Bayesian inference (Xing et al., 2011), optimizing cost function and motion vector (Mackowiak, 2013), snake (Lefèvre et al., 2000; Lefèvre and Vincent, 2004), color-based PF, meanshift, Kalman-meanshift (Nummiaro et al., 2003), optimal path search using dynamic programming in graph (Pallavi et al., 2008), graph and particle swarm optimization (PSO) (Manafifard et al., 2015), PF (Dearden et al., 2006; Choi et al., 2004; Ok et al., 2002; Chai et al., 2011; Davis, 2008), hierarchical PF (Yang et al., 2005; Wang et al., 2008), support vector regression PF (Zhu et al., 2006), PF incorporated by MCMC (Zhang et al., 2008), boosted interactively distributed PF (BIDPF) (Wu et al., 2008), combined probabilistic support vector classification (PSVC) with support vector regression PF (Zhu et al., 2009), local TSV and simple blob tracking (Sato and Aggarwal, 2005), MHT (Beetz et al., 2006) and spatial similarity matrix (SSM) (Duh et al., 2013), have been applied for player tracking in videos captured by moving camera including broadcast streams.

Different tracking schemes, such as match matrix (Martín and Martínez, 2013), KF (Xu et al., 2004a, Ren et al., 2009; Misu et al., 2004), minimizing occupancy-based energy function (Khan and Shah, 2009), JPDA-KF (Abbott and Williams, 2007), tracking-by-detection using similarity measure (D'Orazio et al., 2009), PF (Du and Piater, 2007), improved PF (Sentioscope) (Baysal and Duygulu, 2016), minimal path searching in graph (Figueroa et al., 2006), graph (Sullivan and Carlsson, 2006), matching SIFT features (Li and Flierl, 2012), PF for each view and K-partite graphs for fusing multiple views (Hamid et al., 2010), nearest neighbor (NN) based on distance, label, area and fusing information of cameras by averaging using homography (Iwase and Saito, 2003), NN for 2D tracking and fusing information of cameras using fundamental matrix for 3D tracking (Iwase and Saito, 2002) have also been applied for player tracking in videos captured by multiple stationary cameras. Moreover, trackers such as, interacting multiple models (IMM) combined with meanshift (Zhong et al., 2006), snake (Vandenbroucke et al., 1997a), PF and KF (Needham and Boyle, 2001), motion detection by checking pixel distribution based on entropy on different directions (Mentzelopoulos et al., 2012) and hierarchical PF (Vos and Brink, 2009), have been applied for player tracking in videos captured by single stationary camera.

Interest point tracking by point distribution models (2D tracking), KF and association (Hayet et al., 2005) and Rao-Blackwellized resampling PF (RBRPF) (Hoyningen-Huene and Beetz, 2009) have also been applied for player tracking captured using multiple moving cameras. In Misu et al. (2009), mixed cameras were applied. For this purpose, player trajectories and identities from a fixed wide angle camera and a motion-controlled camera were fused for handling occlusions. It also required a special apparatus for measuring and controlling camera parameters.

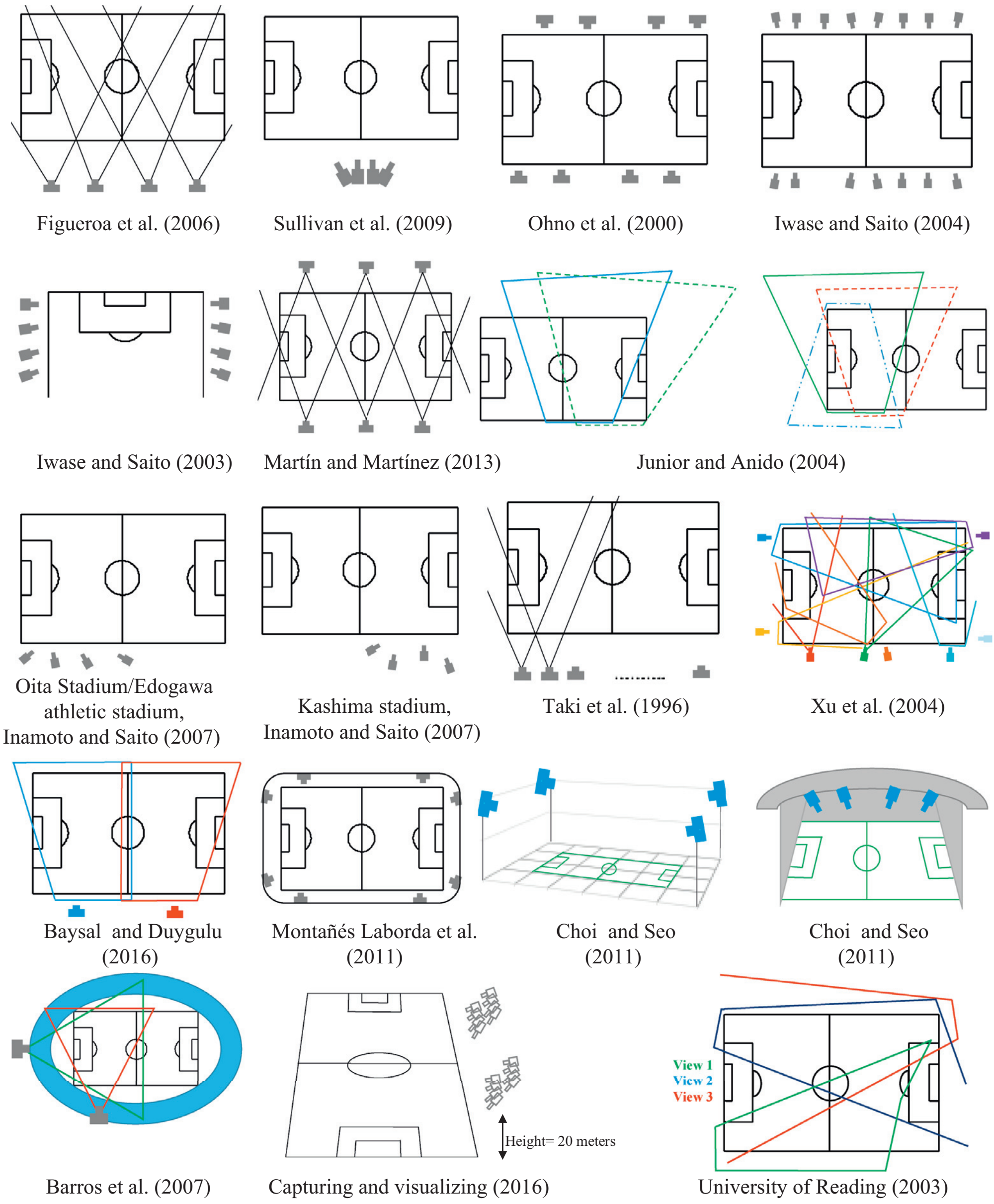
Generally, much work has been done on broadcast videos, and multiple moving cameras has got few attentions. The main camera used to capture the main area of activities in broadcast videos is often fixed in location near the center line of the field. It is also free to rotate and zoom; however, it usually cannot roll (i.e. rotate about its direction of view). Although broadcast videos are edited by the director through the display of close ups or replays, most of the frames are long shots representing a global view of the field. In general, monocular approaches are affected by different problems, such as small field of view (FOV), lack of 3D information, low resolution and complex occlusions. The broadcast cameras also partly cover the playfield, and all players are not visible all the time.

An alternative approach to improve soccer video analysis is to use multiple stationary cameras, which presents a series of advantages. Firstly, ambiguities particularly during crowded, cluttered or occlusion scenes can be reduced (Figueroa et al., 2006), and the analysis results (e.g. player labeling) can be improved (Choi and Seo, 2011). Moreover, the whole playfield can be covered all the time. This makes these setups more suitable for analysis of the game; since the properties of the soccer match, such as predefined numbers of players or their teams, can also be applied (Choi and Seo, 2011; Sullivan and Carlsson, 2006). Furthermore, camera parameters remain fixed (Xu et al., 2004a; Ren et al., 2009), and missed players can be compensated in some views (Iwase and Saito, 2003). Moreover, the best camera for image analysis can be chosen, since the field of views of cameras can be defined (Figueroa et al., 2006; Figueroa et al., 2004). This setting also makes the background subtraction applicable to simplify foreground analysis (Ren et al., 2009). Moreover, the size of the represented player can be big enough to provide information with better quality. 3D information of the scene can also be retrieved, and the accuracy of localization can be improved by the collaboration among multiple views (Xu et al., 2004a; Khan and Shah, 2009), which is extremely required for detecting events such as offside moments (D'Orazio et al., 2009). Despite all these advantages, some limitations of these configurations consist of increasing hardware cost and processing time by increase in number of views (Khan and Shah, 2009), synchronizing different views, considering robustness in configuration and collaboration among multiple views. Moreover, the minimum number of cameras for covering the field depends on factors, such as camera resolution, height and FOV. However, the accuracy may be still not enough in areas of high density where the foot positions of some players are not detected in any frames.

The soccer matches have been captured from cameras at different frame rates (15 frames per second (fps) (D'Orazio et al., 2009; Iwase and Saito, 2004; Iwase and Saito, 2003; Needham and Boyle, 2001; Lefèvre et al., 2000; Yoon et al., 2002; Li and Flierl, 2012; Hashimoto and Ozawa, 2006), 25 fps (Baysal and Duygulu, 2016; Liu et al., 2009; Heydari and Moghadam, 2012; Zhang et al., 2008; Mazzeo et al., 2008; Najafzadeh et al., 2015; D'Orazio et al., 2007; Kayumbi et al., 2008; Tong et al., 2011; Nunez et al., 2008; Tabii and Thami, 2009; Ekin et al., 2003), 29 fps (Nunez et al., 2008), 30 fps (Choi and Seo, 2011; Khan and Shah, 2009; Kasuya et al., 2008; Utsumi et al., 2002; Sato and Aggarwal, 2005; Ekin et al., 2003; Itoh et al., 2012)). Although high resolution images (e.g. 1920 × 1080 (D'Orazio et al., 2009; Mazzeo et al., 2008; D'Orazio et al., 2007), 1388 × 1036 (Montañés Laborda et al., 2011)) are beneficial to the tracking and occlusion resolution, medium resolution images (e.g. 720 × 576 (Choi and Seo, 2011; Joo and Chellappa, 2007; Liu et al., 2009; Maćkowiak et al., 2010; Sato and Aggarwal, 2005; Hoyningen-Huene and Beetz, 2009; Manafifard et al., 2015; Wu et al., 2008; Najafzadeh et al., 2015; Tong et al., 2011; Herrmann et al., 2014; Mackowiak and Konieczny, 2012; Manafifard et al., 2016), 720 × 480 (Choi and Seo, 2011; Iwase and Saito, 2004; Iwase and Saito, 2003; Inamoto and Saito, 2007; Chiang et al., 2009; Sato and Aggarwal, 2005; Wang et al., 2008; Hashimoto and Ozawa, 2006)) or low resolution images (e.g. 352 × 288 (Pallavi et al., 2008; Yang et al., 2005; Nunez et al., 2008; Tabii and Thami, 2009; Ekin et al., 2003), 320 × 240 (Utsumi et al., 2002; Mochizuki et al., 2009; Watanabe et al., 2004)) were often used due to the effect of increasing image resolution on the computational time (Khan and Shah, 2009). Moreover, the soccer match has been captured by different number of cameras (one (Sullivan and Carlsson, 2006; Spagnolo et al., 2007), two (Kasuya et al., 2008; Hoyningen-Huene and Beetz, 2009; Misu et al., 2009; Du et al., 2006), three (Mentzelopoulos et al., 2012; Enomoto and Saito, 2009; Du and Piater, 2007; Li and Flierl, 2012), four (Choi and Seo, 2011; Figueroa et al., 2006; Inamoto and Saito, 2007; Abbott and Williams, 2007; Sullivan et al., 2009; Hayet et al., 2005; Figueroa et al., 2004), six (D'Orazio et al., 2009; Martín and Martínez, 2013; Baysal and Duygulu, 2016; Kayumbi et al., 2008; Ben Shitrit et al., 2014), eight (Xu et al., 2004a; Choi and Seo, 2011; Ren et al., 2009; Khan and Shah, 2009; Montañés Laborda et al., 2011; Iwase and Saito, 2003; Xu et al., 2004b), fifteen (Iwase and Saito, 2004)). However, the distance between cameras in multi-view configurations and their camera height were rarely reported in the literature. In Enomoto and Saito (2009), the distance between the first and second camera was 10 m, and the distance between the second and third camera was 20 m. In a different multi-view configuration (Figueroa et al., 2006), four cameras were placed at the height of 20 m and distance of 40 m from the field line on one side of the pitch. In Beetz et al. (2007), the broadcast video was captured by the main camera at the height of 8–22 m in the stands near the center line of the field. Moreover, position and distance between cameras can be chosen to cover different parts of the field with overlapping regions (Figueroa et al., 2006). The position of the cameras has also been chosen by the layout of the stadium to achieve the best coverage of the field and the best resolution of each area (Xu et al., 2004a, Ren et al., 2009).

A number of different multi-camera configurations are shown in Fig. 3. One typical configuration is multiple fixed cameras located on one side of the field (Inamoto and Saito, 2007). These cameras can be fixed at the highest location of the stadium (Figueroa et al., 2006; Barros et al., 2007), along the touchline (Taki et al., 1996) or on a tripod (Sullivan and Carlsson, 2006; Sullivan et al., 2009), and there may exist extra overlapping regions (Figueroa et al., 2006). Moreover, the optical axis of each camera can be perpendicular to the touchline (Taki et al., 1996) or not. Another typical approach is locating stationary cameras on both sides or around the field (Xu et al., 2004a; Choi and Seo, 2011; Ren et al., 2009; Khan and Shah, 2009; D'Orazio et al., 2009; Iwase and Saito, 2004; Martín and Martínez, 2013; Montañés Laborda et al., 2011; Junior and Anido, 2004). In Montañés Laborda et al. (2011), eight static cameras were positioned on the roof around the stadium, and each side of the field and each goalpost area were covered by two cameras. The system in Iwase and Saito (2003) used eight static cameras focusing on the penalty area at both sides of the field. In Misu et al. (2009), two types of cameras were used. One was a fixed wide angle camera for player tracking, and the other was a motion controlled camera with zoom lens for player recognition during occlusions whose parameters were measured by rotary encoders. In Baysal and Duygulu (2016), one camera captured the left, and the other captured the right half of the field. A narrow portion of the field along the midline was also common in both cameras. The TRACAB player tracking system (Capturing and visualizing large scale human action, 2016) consisted of two small multi-camera units, and every inch of the pitch was covered by at least two cameras. Moreover, each camera could achieve a higher resolution in crowded part of the pitch and occlusion situations.

One important issue in multi-view configuration is the association (Section 6.10) and fusion across multiple simultaneous views. Association concerns with assigning each measurement to its associated player, and fusion concerns with fusing measurements from multiple views. Therefore, good observations in some views can compensate for poor observations in other views, and different tracks can be fused across different cameras (Martín and Martínez, 2013). In Martín and Martínez (2013), thresholding the score defined by blobs' distances was applied for deciding which blobs should be fused. In multiple camera techniques, the best view can be chosen, or the results from multiple views can be fused. The best view can be picked depending on the distance of the player to the center of the camera FOV (Figueroa et al., 2006). In Choi and Seo (2011), a player on 3D field was won by the camera with best resolution. In Xu et al. (2004a), player estimation was shifted to the most accurate measurement captured by the closest camera. Since the degree of confidence for all views is not the same, fusion can be performed by weighting results from each view. In Khan and Shah (2009), higher fusion weights were assigned to views with less clutter. The method by Choi and Seo (2011) selected the players' class regarding the maximum weighted sum of probabilities from the associated blobs. The association and fusion of information from multiple views can be image-based. Accordingly, the geometrical relationship between cameras was made by homography (Sullivan and Carlsson, 2006; Iwase and Saito, 2003) or fundamental matrix (Iwase and Saito, 2002) while the result using homography outperformed. In order to fuse data, the coordinates of the players from multiple views were first averaged by Iwase and Saito (2003), and then the second average between close coordinates to the first average was considered as the estimated player location. In Iwase and Saito (2002), the occluded player location was estimated by calculating an intersection of two epipolar lines corresponding to player locations from non-occluded farthest cam-



Figueroa et al. (2006)

Sullivan et al. (2009)

Ohno et al. (2000)

Iwase and Saito (2004)

Iwase and Saito (2003)

Martín and Martínez (2013)

Junior and Anido (2004)

Oita Stadium/Edogawa athletic stadium, Inamoto and Saito (2007)

Kashima stadium, Inamoto and Saito (2007)

Taki et al. (1996)

Xu et al. (2004)

Baysal and Duygulu (2016)

Montañés Laborda et al. (2011)

Choi and Seo (2011)

Choi and Seo (2011)

Barros et al. (2007)

Capturing and visualizing (2016)

University of Reading (2003)

**Fig. 3.** Different camera configurations.

eras. A different approach was performing association and fusion in real world (i.e., model-based) (Poppe et al., 2010). Following the projection of player bounding boxes onto the field, nearby feet can be assigned to the same player (Junior and Anido, 2004). In Iwase and Saito (2004), each projected player position onto the field was assigned to the closest estimated position by Kalman, and the centroid of the projected positions assigned to each player was calculated. Then, the estimated positions were updated with the centroids. The method by Choi and Seo (2011) performed association regarding the distance between feet projection of each view and the feet projection of the view with the best resolution on the field. Another solution was associating trajectories' segments from multiple views on the field model regarding similarity between them and performing data fusion by average of the associated pair (Kayumbi et al., 2008). The method by Hamid et al. (2010) modeled fusion problem using K-partite graphs in the real world, and observations were fused by finding minimum weight K-length cycles in the K-partite graphs. As a result, the fusion outperformed naively fusing information. In Hayet et al. (2005), the best association was found by combinatorial optimization which minimized the Mahalanobis distance between the ground player position and the predicted player position. Then, the measurements from different cameras were fused regarding the uncertainty in the locations of the tracked points and features for computing homography.

## 3. Playfield detection

Playfield detection is often the first stage and plays a primary role in soccer video content analysis. It not only detects the playfield region to reduce noise from non-playfield areas, but also identifies foreground objects (e.g. players, field lines and circles) from background by filtering out the grass pixels and preserving the objects in the green playfield. Therefore, it reduces the amount of pixels to be processed and noise for simplifying the player detection and tracking stages. Since the soccer field is predominantly green in long and medium shots, most researchers have tried to detect the playfield by detecting the dominant color, filling the interior of the largest connected component and defining its

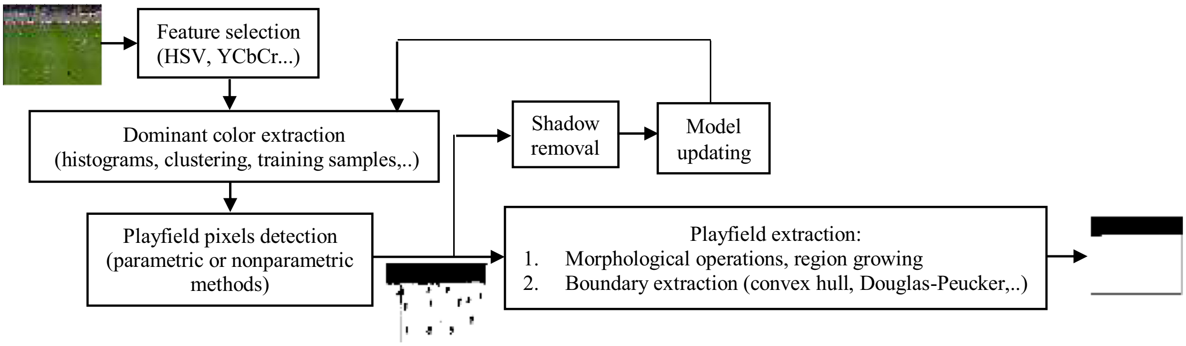

Feature selection (HSV, YCbCr...)
Dominant color extraction (histograms, clustering, training samples,..)
Playfield pixels detection (parametric or nonparametric methods)
Shadow removal
Model updating
Playfield extraction:
1. Morphological operations, region growing
2. Boundary extraction (convex hull, Douglas-Peucker,..)

**Fig. 4.** The flowchart of playfield detection component.

convex hull. The color of the playfield may vary from stadium to stadium or within one stadium depending on time of the day, weather and lighting conditions, shadows and the viewing angle. Therefore, accurate playfield segmentation cannot be achieved by learning playfield color and holding it fixed without updating its statistics during the game. There are also noises due to green patches in the players' shirts, crowd and advertisement boards, and each light source may produce a distinct shadow at the base of a player. The main steps of playfield detection (i.e. feature selection, dominant color extraction, playfield pixels detection, playfield extraction, shadow removal and model updating) are shown in Fig. 4. However, some of these steps (i.e., shadow removal, model updating) are neglected in most of the previous works.

Most researchers have tried to detect the playfield regarding the existence of a single dominant color in the far view shots. Accordingly, some researchers have tried to utilize image features independent of illumination by transforming from the RGB space to HIS (Davis, 2008; Spagnolo et al., 2007; Le Troter et al., 2004), YCbCr (Heydari and Moghadam, 2012), normalized rgb (Barnard and Odobez, 2004) or YIQ (Pallavi et al., 2008).

In most previous approaches, field color was first learnt via supervised approaches, such as learning field color from training sets (Vandenbroucke et al., 1998) or unsupervised approaches, such as clustering (Barceló et al., 2005). For the supervised category, there exist some works in which predefined thresholds were set experimentally for detecting the playfield pixels (Pallavi et al., 2008; Utsumi et al., 2002; Watanabe et al., 2004; Kangarloo and Kabir, 2005; Hoernig et al., 2015). The main drawback of this approach was the dependence of the thresholds to the grass color which differs regarding the stadium and lighting conditions. Several methods, such as mean, dynamic sliding window and k-means, were introduced in Davis (2008) for selecting the thresholds. The playfield color can also be represented by color peak value of the histogram (Choi et al., 2004; Li and Lu, 2007; Li et al., 2009; Ngo et al., 2010). Some methods calculated the mean value of each color component around their respective histogram peaks (Tabii and Thami, 2009; Ekin et al., 2003) or defined an interval around the histogram peak (Vandenbroucke et al., 1997a, b; Nunez et al., 2008; Vandenbroucke et al., 1998).

Approaches for playfield pixels detection can be divided into two main approaches, namely, non-parametric and parametric methods. Gaussian mixture models (GMM) (Sullivan and Carlsson, 2006; Zhu et al., 2006; Manafifard et al., 2015; Davis, 2008; Duh et al., 2013; Manafifard et al., 2016; Barnard and Odobez, 2004; Kangarloo and Kabir, 2005; Ngo et al., 2010; Gedikli et al., 2007; Jiang et al., 2004; Bu et al., 2011; Wang et al., 2004) and color histogram-based method (Liu et al., 2009; Dearden et al., 2006; Huang et al., 2007; Khatoonabadi and Rahmati, 2009; Yoon et al., 2002; Seo et al., 1997; Tong et al., 2011; Sun and Liu, 2009; Assfalg et al., 2003; Xu et al., 2001) were the most widely used parametric and non-parametric techniques, respectively. Following dominant color extraction, playfield pixels could be detected either by thresholding the distance between the color of testing pixel and dominant color (Barceló et al., 2005; Ekin et al., 2003; Li et al., 2009; Sun and Liu, 2009; Tran et al., 2012) or calculating the distance from the field color class in the case of learning dominant color by GMM (Davis, 2008; Gedikli et al., 2007). In Davis (2008), the grass region was detected using background subtraction, and then playfield pixels were classified using Mahalanobis distance from the playfield model represented by Eigen model and GMM. As a result, both models achieved similar levels of success; however, GMM had higher computational complexity. In addition, playfield detection using GMM outperformed that of histogram and hidden Markov models (HMMs) in Jiang et al. (2004). Alternatively, thresholding (Yoon et al., 2002) or clustering (Hung et al., 2011) based on histogram peaks was applied for detecting playfield pixels. An improved generalized Lloyd algorithm (GLA) (Yu et al., 2007) and multilayer perceptron neural network (MLP) classifier (Kangarloo and Kabir, 2004) were also applied for playfield detection.

A combination of features was used by few algorithms for playfield detection. For instance, the method by Xing et al. (2011) used multiple cues including color, motion, and shape for playfield segmentation in two online and offline steps. The method by Sullivan and Carlsson (2006) applied gradient with GMM instead of color features, and the method by Bai et al. (2011) fused two binary playfield detection results from color ratio and local entropy thresholding. In Kangarloo and Kabir (2004), color (hue) and texture (wavelet) features along with MLP classifier were applied for playfield detection. Another approach by Xu et al. (2004a) defined the pitch mask as the intersection of color-based and geometry-based masks. The geometric constraint, namely, area of coverage based on homography was also applied by Ngo et al. (2010). In a different approach by Naemura et al. (2000), geodesic binary reconstruction was combined with segmentation by HIS (Hue, Intensity, Saturation) color histogram. In Le Troter et al. (2004), color in HLS (Hue, Luminance, Saturation) space and spatial coherence were used for detecting the playfield. In Kangarloo and Kabir (2005), hue and texture were used as features to model the grass as mixture of Gaussians (MOG), and the grass was segmented using multi-layer perceptron and Bayes theory.

In order to deal with shadows, intrinsic image (Liu et al., 2011), geometric constraints of multi-camera setup (Kasuya et al., 2008; Hamid et al., 2010) and skeletonisation algorithm (Renno et al., 2004) have been presented. Field color should also be adapted during the game. In Sato and Aggarwal (2005), the mean value of randomly selected pixels was computed as the field color, and the final field color was obtained by weighted sum of the color of sample pixels in the previous and current frame. The method by Wang et al. (2004) used the extracted playfield to re-estimate the GMM, and an incremental training was applied by

Duh et al. (2013) for generating an adaptive Gaussian field model. Finally, the playfield mask was often obtained by applying morphological operations on the resulting binary image from the playfield pixels detection, and the field was extracted by the convex hull of the green field (Dearden et al., 2006; Davis, 2008; Manafifard et al., 2016; Hoernig et al., 2015; Yu et al., 2007) or boundary-following algorithm (Seo et al., 1997). The region growing method has also been employed by Zhu et al. (2006) and Jiang et al. (2004) to connect playfield pixels into regions.

## 4. Player detection

Player detection (i.e. player region recognition) is another crucial stage for soccer video analysis. Player detection, as a low level image processing task, provides measurements which can be used for initializing the subsequent tracking step (e.g. KF or graph tracking). The player detection quality is of great consequence due to its direct impact on the player tracking step. Following player detection, each player can be approximated by a bounding box or an elliptic blob, and the player location in the image can be determined by the position of the player feet. Accordingly, the midpoint of the base of a player bounding box was often taken as the feet position.

Some authors have tried to perform the coarse player detection using connected component analysis after isolating grass and spectator regions and applying morphological operations. However, some noisy regions may remain. In Davis (2008), a feature space representing player model was created from several features (i.e., area, convex area, solidity, major and minor axis, orientation, ellipsity, extent) using a training set. Then, players were detected using Mahalanobis distance to the player model. The method by Yoon et al. (2002) thresholded size, compactness, ratio of vertical to horizontal length and color distribution to separate players' regions while the gray level and the Hough value were thresholded to separate lines in the remaining regions. The method by Khatoonabadi and Rahmati (2009) performed the grass field extraction step with new parameters whenever the player detection failed, and then the region-based detection algorithm was applied. Other approaches used for player detection include motion-based player detection by background subtraction (Xu et al., 2004a, Choi and Seo, 2011; Khan and Shah, 2009; D'Orazio et al., 2009; Martín and Martínez, 2013; Montañés Laborda et al., 2011; Junior and Anido, 2004; Figueroa et al., 2006; Sullivan and Carlsson, 2006; Iwase and Saito, 2003; Inamoto and Saito, 2007; Kasuya et al., 2008; Misu et al., 2004; Hamid et al., 2010; Sullivan et al., 2009; Joo and Chellappa, 2007; Vos and Brink, 2009; D'Orazio et al., 2007; Kayumbi et al., 2008; Spagnolo et al., 2007; Naidoo and Tapamo, 2006; Marchesotti et al., 2004), lazy background subtraction and connected components analysis to limit the search area (Abbott and Williams, 2007), shape-based player detection (Huang et al., 2007), texture-based player detection (Mochizuki et al., 2009), edge-based player detection (Tran et al., 2012), classification schemes, such as support vector machines (SVM), neural networks, adaboost, linear discriminant analysis (LDA) and nearest neighbor (NN) (Baysal and Duygulu, 2016; Liu et al., 2009; Xing et al., 2011; Maćkowiak et al., 2010; Heydari and Moghadam, 2012; Zhu et al., 2006; Manafifard et al., 2015; Davis, 2008; Tong et al., 2011; Gerke et al., 2013; Schlipsing et al., 2014), clustering schemes (e.g. k-means) (Kim et al., 2003; Nunez et al., 2008), pixel-based player detection by classifying each pixel using its color components (Needham and Boyle, 2001; Vandenbroucke et al., 2003; Utsumi et al., 2002; Vandenbroucke et al., 1997b, Naidoo and Tapamo, 2006), template-based player detection by classifying each window in the image into a player or non-player (Heydari and Moghadam, 2012; Zhu et al., 2006; Sun and Liu, 2009; Ekin and Tekalp, 2003) and modified histogram backprojection (Kawashima et al., 1994). Since accurate player detection by background subtraction was difficult in the case of moving cameras, camera motion was compensated through spatio-temporal operators, image to model registration (Intille and Bobick, 1995a, b) or mosaic construction (Barceló et al., 2005; Bebie and Bieri, 1998) before applying background subtraction in some previous works. The main limitation was the need for additional computations relating to the registration and playfield marks removal. The main problem of player detection based on uniform color was choosing appropriate color features which vary for each game. Most authors determined the color spaces more appropriate for their specific detection problem (Needham and Boyle, 2001; Kim et al., 2003; Nunez et al., 2008); however, a hybrid color space was learnt offline by Vandenbroucke et al. (1997a, 2003) to select a set of color components. In clustering schemes, choosing the number of clusters was a challenging issue. Some authors have also tried to segment players in a probabilistic manner in which a likelihood or confidence map was formed (Khan and Shah, 2009; Hamid et al., 2010; Needham and Boyle, 2001; Beetz et al., 2007; Gedikli et al., 2007; Yao et al., 2010). In Yao's method (Yao et al., 2010), the confidence map was generated from the output of a Hough forest trained for athlete segmentation. In Khan and Shah (2009), the foreground likelihood maps obtained by background subtraction from all views were fused to produce a synergy map representing the likelihood of players occupying the locations on the scene. In the blob-guided PSO method proposed by the authors of this paper (Manafifard et al., 2016), a two-step blob detection step was combined with an efficient search mechanism based on PSO for player detection. Partial occlusions were handled in this approach, while most of the previous player detectors (e.g. Adaboost, SVM, neural network and background subtraction) relied on binary classification without locating players in partially occluding blobs.

Different features, such as color, edge, texture, shape and motion, have been used for player detection. Since teams are dressed to be distinguished easily, the most informative feature is color which is also robust against scaling, rotation, partial occlusion, and non-rigid deformation. However, it suffers from illumination changes or the presence of confusing colors in the background. Color feature is not also reliable across the large scenes without adaptation. Furthermore, the small number of pixels representing a distant player prevents building an accurate color player model. It is also really important to select a suitable color space. However, appropriate color space extremely depends on player uniform that changes from one game to another. Alternatively, players have locally strong edges and edge features are more robust against illumination changes, but they are sensitive to clutter. In order to take advantage of gradient features, histogram of gradients (HOG) has been applied along with some classifiers (e.g. SVM) (Baysal and Duygulu, 2016; Mackowiak, 2013). Considering the limitation of using a single feature, some authors have also tried to combine color with other features, such as local edge property (Utsumi et al., 2002), gradient (Naemura et al., 2000; Manafifard et al., 2016; Gerke et al., 2013), motion (Schlipsing et al., 2014) and entropy (Mentzelopoulos et al., 2012). In Gerke et al. (2013), HOG and three color features (block-based HSV color histograms, color spatiograms, color and edge directivity descriptors) were used. As a result, detection was improved using the classifier fusion with HOG. Viola and Jones' framework (Viola and Jones, 2004) was also improved for detecting players by adding gradient information in Xing et al. (2011). As a result, it outperformed the HOG boosted human detector. To sum up, color is the most informative feature which outperforms intensity or gray level features. In particular, it is required for isolating partially occluding players. One natural way to improve performance is combining color with other features. Accordingly, the gradient is one of the most appropriate features to be combined with color for discriminating players and lines, since the gradient directions of points on line segments are mostly similar. Moreover, features such as size, area, coverage ratio, extent, orientation, convex area, solidity, ellipsity and compactness should be applied as auxiliary features**.** Player detection can also be performed using entropy, texture and Haar features according to the distinctive structure of the players in the green field; however, these features would be more effective along with color cues. Alternatively, temporal variables and motion features within the background subtraction approach were the most widely used features in the case of static cameras. Also, shape features were rarely used for player detection due to the small player size and variable player shape during the game (e.g. fallen player).

The main works published in player detection are summarized in Table 1. In the columns, the applied methods, feature descriptors, camera status (moving or static camera), number of test frames, strengths and weaknesses are presented. Since the player detection provides observations for the player tracking step, the number of frames for player tracking is also included in the fifth column. Moreover, only the main features are presented in the third column. Since constraints (e.g. size and area) were usually used by most researchers, they are only denoted as descriptors when they were the main applied features. Moreover, only applying information on uniform color is meant by color descriptor, since information on playfield color was used in most previous works. The size descriptor is also described in general, which can include the bounding box length and width, vertical to horizontal length or even the area.

One weakness of most player detectors is relying on binary classification without locating players in partially occluding blobs (NI POC in Table 1). Furthermore, the main limitation of simple background subtraction is that the non-stationary cameras in broadcast videos hinder its use (UM in Table 1) due to the combination of players and camera motions. Moreover, missed detections can result from players standing still or moving very slowly for a long time. The main limitation of background subtraction using camera motion compensation is also the need for additional computations relating to the registration and playfield marks removal and the calibration inaccuracy due to the low textured field. The color of player uniform can also be used for isolating partially occluding competitors; however, it requires to be learnt for each game (LCG in Table 1). Moreover, the player pixels may be misclassified when the colors of the legs or socks of players from one team are similar to the ones from opponent suit or lines. Discrimination of small white players and lines, particularly occluding ones (SW in Table 1), is also challenging using color cues. In this case, the image gradient is beneficial, which also increases robustness to illumination changes. The main problem with the pixel-based player detection is also the fragmentation of a player (particularly the players' legs) into multiple regions due to differences in the color of shorts, jerseys, and socks. However, different poses of players are handled (CDP in Table 1) by background subtraction and pixel-based classification, which are more challenging for template-based detectors (e.g. classifiers). Moreover, player detection by classifiers often suffered from a time consuming training phase (L-TR in Table 1). Few attempts have also been made at segmenting players in an unsupervised manner by using clustering schemes. However, choosing the number of clusters corresponding to player clothes colors and choosing the initial cluster center are still challenging problems. Although results are improved by classifiers compared to unsupervised schemes, they strongly depend on the number and quality of training set and applied features. To sum up, background subtraction (D'Orazio et al., 2009 ) in the case of static cameras and classifiers (Schlipsing et al., 2014) in the case of moving cameras are the two most effective detectors in the literature; however, they relied on binary classification. In order to solve this problem, a blob-guided PSO method (Manafifard et al., 2016) was proposed by the authors of this paper.

## 5. Appearance modeling and player labeling

Following player detection, an optional step for semantic analysis of the game is team discrimination or player classification into the five classes corresponding to the two teams, two goalkeepers and referee, which is known as player labeling. Since distinguishing between players of the two teams can be performed using distinctive color patterns in uniforms, labeling has been simultaneously performed with color-based player detection (e.g. Vandenbroucke et al., 2003). Moreover, labeling was either solved independently as an input to the tracking step (e.g. graph tracking (Figueroa et al., 2006)), or it was performed during the tracking step (e.g. PF (Zhang et al., 2008; Yao et al., 2010)). In the case of simultaneous labeling and tracking schemes, candidate regions (e.g. templates) in the input image were compared to the player region from the previous frame to pick up the most similar position as the player location in the current frame. Even when labeling was performed independent of tracking, its performance could be improved after the tracking step (Liu et al., 2009). As shown in Fig. 5, color-based player detection, player labeling and some trackers are performed through the comparison of the player model (appearance model) and the player candidates by similarity measures. Therefore, defining an appropriate player model and updating it through the sequence are essential (Jahandide et al., 2012). Although primitive geometric shapes (e.g. rectangle, ellipse) were the most widely used method for player shape representation (D'Orazio et al., 2009; Xing et al., 2011; Khatoonabadi and Rahmati, 2009; Nummiaro et al., 2003), there have been few works for point-based player tracking (Hayet et al., 2005; Gabriel et al., 2005), contour-based player tracking (Vandenbroucke et al., 1997a, Lefèvre et al., 2000) and skeletal models (Huang et al., 2007). Moreover, player appearance has been represented by mixture of Gaussians (Montañés Laborda et al., 2011; Davis, 2008), histograms (Xu et al., 2004a; Baysal and Duygulu, 2016; Hamid et al., 2010; Chiang et al., 2009; Khatoonabadi and Rahmati, 2009; Nummiaro et al., 2003; Duh et al., 2013; Najafzadeh et al., 2015; Du et al., 2006), weighted histogram (Manafifard et al., 2016), color spatiograms (Gerke et al., 2013; Schlipsing et al., 2014), PCA (Schlipsing et al., 2014), color and edge directivity descriptor (CEDD) (Gerke et al., 2013), polar appearance representation (Kang et al., 2004), mean (Hashimoto and Ozawa, 2006), vertical distribution (Figueroa et al., 2006; Barceló et al., 2005; Yoon et al., 2002; Seo et al., 1997; Figueroa et al., 2004), combination of color and edge features (Yang et al., 2005). In order to account for spatial information, spatial-color mixture of Gaussians (SMOG) was introduced by Wang et al. (2006) where both the color and spatial information were utilized. As a result, it was more discriminative than color histograms. In our previous work (Manafifard et al., 2016), a player template was partitioned into some blocks to account for spatial information, and the weighted histograms were computed for player labeling. Appearance variations have also been handled through adaptive methods by adjusting size (Qian et al., 2007) and color (Montañés Laborda et al., 2011; Zheng and Xue, 2009; Nummiaro et al., 2003; Wang et al., 2006).

Generally, player labeling can be done independently or along with the detection and tracking steps. After the player model definition, each player was assigned to one of the relative classes using similarity measures such as color distances (D'Orazio et al., 2009; Hamid et al., 2010; Davis, 2008; Hashimoto and Ozawa, 2006; Manafifard et al., 2016; Naidoo and Tapamo, 2006), histogram intersection (Yoon et al., 2002) or classification schemes (Choi and Seo, 2011; Baysal and Duygulu, 2016; Kasuya et al., 2008; Tran et al., 2012). The method by Sullivan and Carlsson (2006) solved the problem in a probabilistic manner, and the player blob was classified as the category with the maximal likelihood. Since the colors of players' uniforms varied for each game, it was beneficial

**Table 1**
A review of player detection (M/S: moving or static camera, Des (C/G/H/M/E/S/Eg/GL/CO/I/T/SH/CR/TV/O/EX/SO/EL/CA): Descriptor (color, gradient, Haar feature, motion, entropy, size, edge, gray level, compactness, intensity, texture, shape, coverage ratio, temporal variables, orientation, extent, solidity, ellipsity, convex area), -: unclear or unmentioned, BS: background subtraction, NI POC: not isolating players in partial occlusion using a single frame, UM: unsuitable for moving cameras, S-IL: sensitivity of RGB color space to illumination changes, L-TR: requiring large training set or difficult training, OTS: optimizing training speed, SW: sensitivity to white players, HC: using hard constraints or thresholds, CDP: considering different poses of players, FG: fragmenting a player into multiple regions, LCG: the need for learning the color of uniforms for each game, CNC: choosing the number of clusters and initial cluster centers is challenging, F-IN: fusing information across multiple cameras, UB: updating background model, AF: affected, Outp: outperformed).

| Ref. | Method | Des | M/S | Frames | Strengths | Weaknesses |
|---|---|---|---|---|---|---|
| Liu et al. (2009) | Gentle Adaboost | H | M | Hundreds of frames from two videos (50–100 frames interval between adjacent frames) for detection, 100 (first clip) and 250 (two other clips) for tracking | Fast | NI POC, L-TR |
| Macḱowiak et al. (2010) | SVM | G (HOG) | M | 9 sequences with length of 25 to 50 frames | Using G, CDP using 3 SVM | NI POC, L-TR |
| Xing et al. (2011) | Adaboost | G | M | 100 (soccer) and 100 (basketball) for detection, 40 (tracking occluding blobs), five football (2593), five basketball (1572) and hockey (at least 100 frames) sequences for tracking | OTS, fast, Outp HOG boosted detector, using G | NI POC |
| Manafifard et al. (2015) | Adaboost | H | M | 491 | Fast | NI POC, L-TR |
| Zhu et al. (2006) | SVM | C | M | 3599 | - | NI POC, SW, LCG |
| Baysal and Duygulu (2016) | SVM | G (HOG) | S | 60,000 player and 60,000 non-player from over 20 matches (90% for training and 10% for testing) | Using G | NI POC, L-TR |
| Heydari and Moghadam (2012) | Neural network | C, area | M | 1080 | - | NI POC, L-TR, SW, using area, LCG |
| Mackowiak (2013) | SVM and HOG-PCA | G | M | - | Using G, PCA-HOG Outp HOG, CDP using 3 SVM | NI POC, L-TR |
| Beetz et al. (2007) | Likelihood map | C, S, CO, M | M | 344 | Real-time, CDP | NI POC, S-IL, HC, LCG |
| Needham and Boyle (2001) | Pixel-based classification in HIS color space | C | S | Every fifth frame from 835 frames | - | NI POC, SW, HC, LCG |
| Vandenbroucke et al. (2003) | Pixel-based classification in hybrid color space | C | S | 4 (example 1), 6 (example 2) | Isolating POC, using a hybrid color space, CDP | SW, FG, LCG |
| Nunez et al. (2008) | Clustering | C | M | A sequence of minimum 20 frames randomly chosen from each 14 soccer games | - | NI POC, SW, LCG, HC |
| Kim et al. (2003) | K-means clustering | C | M | - | Unsupervised, CDP | NI POC, SW, S-IL, CNC |
| Yoon et al. (2002) | Player detection based on line and grass field removal and some constraints | C, S, GL, CO | M | 40 | - | NI POC, HC, LCG, S-IL, SW |
| Khan and Shah (2009) | Occupancy constraint and BS using GMM | M | S | Parking lot (over 3000), Indoor (-), Basketball (1000), Soccer (1000) | CDP, F-IN, considering occlusions | UM |
| Hamid et al. (2010) | BS using GMM and shadow removal by homography | M, C | S | 60,000 | F-IN, shadow removal, CDP | UM, NI POC |
| Joo and Chellappa (2007) | BS using codebook model and local temporal variance | M | S | 600 (effectiveness validation of soccer), 2500 (quantitative evaluation of soccer), 180 (two non-soccer sequences) | CDP, real-time, UB | UM, NI POC |
| Montañés Laborda et al. (2011) | BS and GMM | M, C | S | - | CDP, real-time | UM, LCG, FG |
| D'Orazio et al. (2009) | BS | M | S | 8 sequences 3000 frames long for detection and tracking group blobs, 2883 (tracking goalkeeper), 15226 (tracking 45 players) | CDP, background modeling without being AF by moving foreground objects, removing shadows, UB, Outp MOG | UM, NI POC |
| Figueroa et al. (2006) | BS using median filter | M | S | 10 min sequence | CDP, UB, detecting most of the still players | UM, NI POC |
| Sullivan and Carlsson (2006) | BS using GMM | M, G | S | More than 1000 | Using G, CDP | UM, NI POC |
| Ren et al. (2009) | BS using GMM | M | S | 2 sequences of 5000 frames | CDP, UB, optimizing speed | UM, NI POC |
| Poppe et al. (2010) | Code-book BS | M, C, I, TV | S | Spagnolo dataset (-) | CDP, UB | UM, NI POC, S-IL |

(*continued on next page*)

**Table 1** (*continued*)

| Ref. | Method | Des | M/S | Frames | Strengths | Weaknesses |
|---|---|---|---|---|---|---|
| Bebie and Bieri (1998) | BS with camera motion compensation | Eg, M | M | - | CDP | NI POC, AF by errors in line removal |
| S.S. Intille and Bobick (1995a, b) | Camera motion compensation and spatio-temporal operators | M, I | M | 270 | CDP | NI POC, using I, AF by registration errors, HC |
| Barceló et al. (2005) | BS with camera motion compensation | M | M | - | CDP | NI POC, AF by errors in line removal |
| Mentzelopoulos et al. (2012) | Entropy difference algorithm, K-means and minimum description length, motion detection | E, C, M | S | 17999 (soccer), 95058 (non-soccer) | Isolating POC, Outp MOG | Complexity, imprecise localization |
| Khatoonabadi and Rahmati (2009) | Multi-thresholding and region-based detection algorithm | C | M | Sequences that were totally eight minutes long, and one fourth of all frames including 15,411 players was evaluated | Outp Yoon et al. (2002) and Ekin et al. (2003) | NI POC, HC, SW, LCG |
| Abbott and Williams (2007) | Lazy BS and connected components analysis | M | S | 100 | CDP, UB, optimizing BS speed | UM, NI POC, new tracks might be problematic |
| Huang et al. (2007) | Reverse Euclidean distance transformation | Sh, C | M | About 300 frames for each three video clips | Extracting a player skeleton | NI POC, SW, HC, S-IL, sensitivity to low resolution |
| Naemura et al. (2000) | Color histogram, watershed, morphological operation and GBR | C, Eg, G | M | About 60 frames for each soccer and baseball scenes | Using G | NI POC, LCG |
| Utsumi et al. (2002) | Fuzzy function to combine color rarity and local edge property | C, Eg | M | 2 video each lasting 30 seconds | - | NI POC, S-IL, LCG, HC, SW |
| Mochizuki et al. (2009) | Classifying image blocks by SVM | T | M | 2 video each lasting 500 seconds with sampling interval of 3 frames | - | NI POC |
| Tran et al. (2012) | Edge pruning | Eg, S | M | (-) frames from 20 soccer matches | Using Eg, CDP | NI POC, HC |
| Schlipsing et al. (2014) | BS, SVM, NN, LDA | C, M | S | 6000 samples, SVM (300,000 samples) | Real-time, OTS, NN Outp SVM and LDA by training from the first few frames and SVM Outp counterparts after a longer training phase | LCG, NI POC |
| Davis (2008) | NN and Mahalanobis distance to the player model | S, area, O, EX, SO, EL, CA | M | 200 | - | NI POC, low precision |
| Manafifard et al. (2016) | Blob-guided particle swarm optimization | C, G, S | M | 8046 (broadcast sequences), every fifty-frames from six clips of the Spagnolo (each clip was 3002 frames long) and every fifty-frames from one subset of the VS-PETS 2003 dataset (2499 frames long) | Using G, isolating POC, reducing the search space, detecting and labeling multiple players simultaneously | LCG, HC |

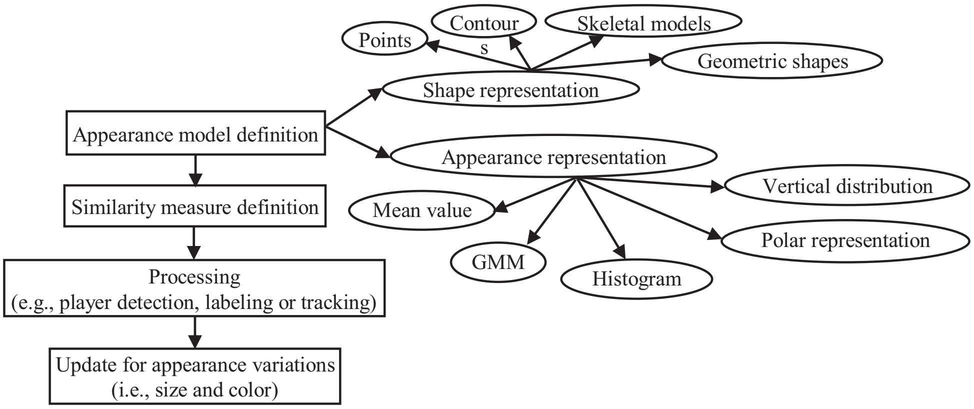


**Fig. 5.** Flow diagram of appearance modeling component.

**Table 2**
A review of player labeling as an independent step (S/U: supervised or unsupervised, S-IL: sensitivity of RGB color space to illumination conditions, IG: ignoring, Cons: considering, SIC: spatial information of color or intensity, F-IN: fusing information across multiple cameras to improve player labeling, I-P: independent from the posture of players, UM: updating model, RT: real-time, OS: optimizing speed, S-P: sensitivity of precision to same colors in the uniforms of different teams, occlusions and player detection result).

| Ref. | S/U | Descriptor | Assignment method | Strengths | Weaknesses |
|---|---|---|---|---|---|
| Liu et al. (2009) | U | GMM in CIE-LUV color space, meta-prototype histogram | Bhattacharya distance | U, I-P | S-P, IG SIC |
| D'Orazio et al. (2009) | U | Normalized color histograms | Manhattan distance | U, I-P | S-P, IG SIC |
| Hashimoto and Ozawa (2006) | S | Mean value | Mahalanobis distance | I-P | S, IG SIC |
| Hamid et al. (2010) | S | Hue and saturation histogram | Bhattacharyya distance | I-P | S, IG SIC |
| Choi and Seo (2011) | S | Camera index, hue-saturation histogram, position of the blob | Support vector machine | F-IN, I-P | S, IG SIC |
| Seo et al. (1997) | S | Vertical distribution of colors in RGB | - | Cons SIC | S, S-IL, not I-P |
| Yoon et al. (2002) | S | Vertical distribution of colors and grays | Histogram intersection | Cons SIC | S, not I-P |
| Barceló et al. (2005) | S | Vertical distribution of colors in RGB | - | Cons SIC | S, S-IL, not I-P |
| Figueroa et al. (2006, 2004) | S | Vertical distribution of intensity | Thresholding | Cons SIC | S, IG color, not I-P, sensitivity to thresholds |
| Xu et al. (2004a) | S | RGB histogram | Histogram intersection | I-P, overcoming distracting background pixels | S, S-IL, IG SIC |
| Khatoonabadi and Rahmati (2009) | S | Histogram | Ratio histogram | Cons SIC | S, not I-P |
| Montañés Laborda et al. (2011) | S | GMM | Distance | UM in RT, OS, I-P | S, IG SIC |
| Tran et al. (2012) | S | RGB histogram | K-means | I-P | S, S-IL, IG SIC |
| Poppe et al. (2010) | S | Color histogram | Bhattacharyya distance | I-P | S, IG SIC |
| Davis (2008) | S | GMM in HSV color space | Mahalanobis distance | I-P | S, IG SIC, using one cluster for modeling each team |
| Manafifard et al. (2016) | S | Weighted color histogram | Bhattacharya distance | Cons SIC, weighted histogram | S, not I-P |

to classify players without the human intervention. Accordingly, few attempts have been made at unsupervised player labeling including improved clustering schemes (Liu et al., 2009; Tong et al., 2011) and basic sequential algorithmic scheme (BSAS) clustering (D'Orazio et al., 2009; Mazzeo et al., 2008; Spagnolo et al., 2007). However, unsupervised precise player shirt determination for unsupervised player labeling at the initial stage of Liu et al. (2009) and Tong et al. (2011) was a challenging task due to the presence of shorts and playfield pixels and player detection errors.

The main methods proposed for the player labeling as an independent step from the player detection and tracking steps are summarized in Table 2. In the second column, the category of the applied approach as supervised or unsupervised is presented. The methods for model description, assigning the players to each team, strengths and weaknesses are also presented in the third, fourth, fifth and sixth column, respectively. Considering spatial information of colors (Cons SIC in Table 2) is an advantage when similar colors exist in the uniforms of different teams. However, labeling will be dependent on the posture of players by considering spatial information. Accordingly, spatial information of colors has been ignored by GMM, mean value and histograms. Since different teams may have a similar histogram, vertical distribution of colors and splitting a player region into several sub-regions (e.g. rectangles) were used to consider spatial information of colors. Moreover, standard RGB color space was applied by some previous works. However, it represents not only the color but also the brightness, and it is sensitive to illumination conditions (S-IL in Table 2). Multiple color patterns can also be modeled by GMM, since using one cluster for modeling each team is inaccurate in the case of different colors in one team uniform. Furthermore, unsupervised labeling methods were very sensitive to same colors in the uniforms of different teams and occlusions. They were also affected by errors in the player detection step, since samples from the player shirt were collected using the detection results.

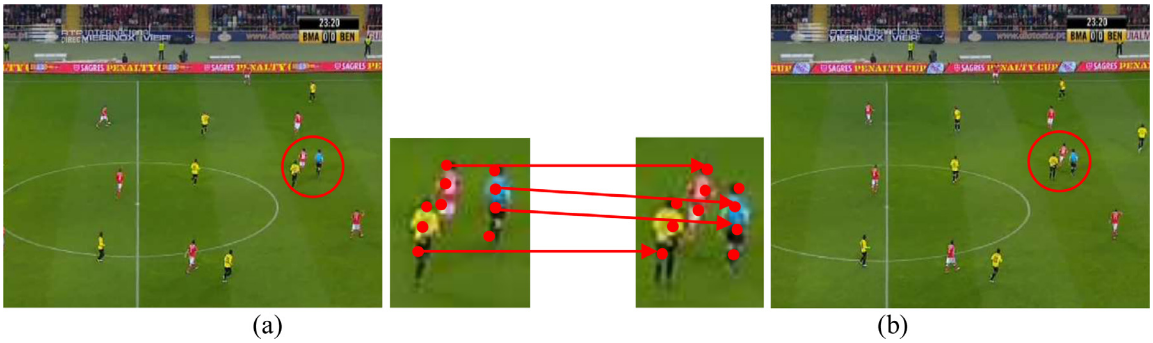


**Fig. 6.** Matching interest points for player tracking.

## 6. Player tracking

Observations corresponding to players (e.g. their position and team affiliation) are delivered by player detection and labeling; however, they lack temporal correspondence. It is therefore necessary to relate observations via a tracker to form a consistent trajectory for each player. Moreover, most of the errors resulting from missed detections, false positives or mislabeling can be resolved by incorporating a tracker. A large number of tracking algorithms have been presented to deal with this topic, such as KF, PF, meanshift, snake, template matching, JPDA, MHT, Markov chain Monte Carlo data association (MCMCDA), etc. Generally, player tracking consists of player state estimation regarding sequential noisy measurements during two main steps, namely, filtering and data association. Filtering concerns unknown state (e.g. player position or size) estimation. However, tracking of multiple players using multiple independent trackers tends to fail for nearby players. Therefore, multiple players tracking involves the problem of data association for jointly tracking of players. Accordingly, data association concerns relating uncertain measurements to each player to work out which measurement is generated by which player. The problem gets more challenging with an increase in the values of false alarm rate, decrease in detection probability and increase in the density of tracks. More details on applied tracking and data association methods are elaborated in the following subsections.

### *6.1. Point tracking*

In point tracking**,** player tracking was formulated as the correspondence of detected player represented by multiple points across the consecutive frames (Yilmaz et al., 2006). Therefore, point tracking involved two main steps, namely, point detection and point correspondence. Few works attempted to track soccer players through point tracking. Since point trackers were suitable for tracking small targets, multiple points were required to track large players (Fig. 6). Accordingly, color interest points extracted by the color version of Harris detector for each player were matched in the current and previous frame by Gabriel et al. (2005). Moreover, each point was characterized by the local appearance of the player along with geometric parameters (point position relative to the estimated center of the region) to avoid matching far points with similar appearances. Another approach using Harris interest points was applied by Hayet et al. (2005) in which point distribution models (PDMs) were constructed by learning the spatial relationships between these points. Then, the players were tracked by matching the points of their PDM to the currently extracted ones. Moreover, variation of interest points was solved by adjustment of the detection scale for local features using homographies. In Li and Flierl (2012), inter-view correlation, inter-frame correlation and motion vector for player tracking were extracted by matching scale-invariant feature transform (SIFT) features. The 3D positions of features were also extracted regarding calibrated cameras, and players were identified by clustering the features. KF and point distribution manifolds were also applied by Mathes and Piater (2006), and the incrementally learnt model combined local appearance and global shape information. Point trackers performed well through partial occlusions. However, it might be hard to detect and match interest points for distant or blurred players.

### *6.2. Contour tracking*

Few works have been presented for contour tracking of soccer players (Vandenbroucke et al., 1997a; Lefèvre et al., 2000) using deformable models (snakes) introduced by Kass et al. [179]. The aim of contour-based trackers was to model the deformable silhouette of players (Fig. 7) more accurately compared to a single shape template. For this purpose, the contour was represented as a set of control points that moved to find player boundary by minimization of an energy function. One limitation of these methods was their sensitivity to parameters, contour initialization, occlusion, non-smooth shape varying process and image resolution. The contour was manually initialized on the first frame in Lefèvre et al. (2000), and it was initialized on the subsequent frames using the position of the snake in the previous frame. Afterwards, the snake deformation based on balloon force was performed, which allowed the contour to expand or shrink like a balloon. However, background pixels, such as lines characterized by high gradient values, misled tracking. The method by Lefèvre and Vincent (2004) applied merge and split steps when two players occluded each other and moved apart. However, these methods were unable to properly handle occlusions.

### *6.3. Graph-based tracking*

Graph representation has been one of the most common solutions to player tracking (Figueroa et al., 2006; Sullivan and Carlsson, 2006; Sullivan et al., 2009; Pallavi et al., 2008; Heydari and Moghadam, 2012; Figueroa et al., 2004; Miura and Kubo, 2008) with capability of handling total or partial occlusions in a merge-split way. Nodes in the graph represent the segmented players while each edge represents the distance between nodes in consecutive frames (Fig. 8). Moreover, each node stores blobs' features (e.g. size, area, center and color). Each edge can also be weighted using the blobs' information (e.g. velocity, orientation, and color) (Figueroa et al., 2006) or the correlation between players corresponding to the edge (Pallavi et al., 2008). Furthermore, the edge can be defined between two overlapped players' blobs (Miura and Kubo, 2008), or edges linking very distant blobs can be removed (Figueroa et al., 2006). Player tracking was usually performed by searching optimal path in the graph. Accordingly, a minimal path searching was used by Figueroa et al. (2006, 2004)

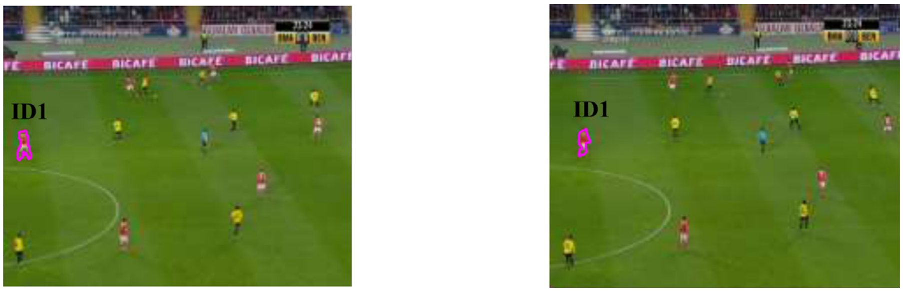


**Fig. 7.** Contour tracking of a soccer player (ID1).

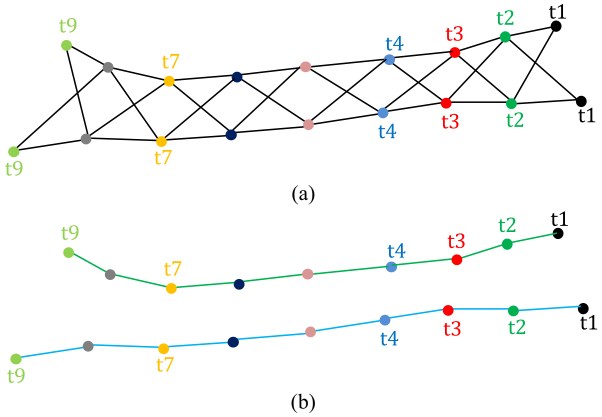


**Fig. 8.** Graph representation. a) Graph representation for two nearby players (tn denotes detected player in $n$-th frame), b) extracted trajectories using the graph representation.

to track isolated players. Optimum path searching was also formulated as a dynamic programming problem by Pallavi et al. (2008). One type of approach was to solve tracking problem as an inference in a Bayesian network to find the most probable set of paths in the graph (Sullivan et al., 2009; Liu et al., 2009; Tong et al., 2011) in which probability for each possible solution could be estimated. In Ben Shitrit et al. (2014), multi-object tracking based on appearance cues was formulated as a multi-commodity network flow problem (MCNF) on a direct acyclic graph (DAG), which was used even when such cues were available at distant intervals. The estimation was performed using linear programming, and a reduced graph was achieved by grouping spatio-temporal locations into tracklets (TMCNF). In our previous work (Manafifard et al., 2015), the specific structure of the graph was exploited to achieve the optimum trajectories using the PSO algorithm.

Generally, occlusions have been solved by exploiting continuity of motion, appearance and relative depth. In Figueroa et al. (2006), the model of each blob before occlusion was searched in the occluding blob to split it. Moreover, prediction was used for solving short occlusions, and the color, trajectories direction and mean velocities of the players were used to maintain the right trajectories. The splitting was also performed in forward and backward directions, and the trajectory was considered the same for players in total occlusions. In our previous work (Manafifard et al., 2015), a combination of neighborhood graph, size of neighboring bounding boxes, blob dimension and color features were employed to predict and split occluding blobs. In Sullivan and Carlsson (2006), occluding players were recovered by a constant velocity motion model when the interpolated positions during the occlusion were foreground. Otherwise, the occluded trajectory was divided into several parts, and the player position at the end of each part was translated to the best location. For occluding teammates, the translation was first found for the closest player to the camera (depth ordering). The method by Figueroa et al. (2006) used the area of blobs to define the number of components. In Miura and Kubo (2008), the range for the number of players in each node was gradually reduced by considering some constraints through several steps. Then, the best combination of player trajectories, which maximized the smoothness measure, was chosen in the heavy occluded scenes; however, the implementation was costly. The player position on the ground was also considered in Figueroa et al. (2006) and Miura and Kubo (2008) to extract player trajectory. Moreover, the location of the lost player could be extrapolated by graph representation, and constraints in the case of fixed cameras, such as fixed number of players, could be applied to match separated trajectories (Sullivan and Carlsson, 2006). Some authors have also applied graph along with other trackers, such as PF (Itoh et al., 2012), due to its capability during occlusions. In graph-based trackers, the number of look forward or backward frames usually increased in the case of long term occlusions, which increased computational cost. Moreover, the splitting of more than three players was more complex, and the prediction results were not satisfactory in the case of abrupt movements.

### *6.4. Template tracking*

The template matching idea is to slide a player template (indicating its appearance) over the image to find a region matching the template. In the tracking context, the search was usually limited around the old player position or predicted player position (Seo et al., 1997). In Taki et al. (1996), each player template was initialized manually from the first frame. Since the tracking procedure sometimes failed for the occluded or falling down player, manual correction was also performed. Other techniques for initializing templates involved using results of the player detection (Yoon et al., 2002) or connected component extraction (Seo et al., 1997), which might be affected by detection errors. One of the first works was presented by Intille and Bobick (1995a, 1995b). Following image to model registration, the player templates were constructed by removing field objects in the closed-world (i.e., space around the player) using intensity, and then template matching was performed. However, it required a large amount of processing, and it faced difficulty in the case of similar players to nearby objects, total occlusions and changing player template during the sequence. In Utsumi et al. (2002), non-occluded players were tracked based on the amount of overlap between the two consecutive frames, and occluded players were tracked by color-based template matching. Another interesting approach has been presented by Khatoonabadi and Rahmati (2009) where large changes due to fast camera movements between consecutive frames were eliminated by back projecting the previous players' positions on the field model into the current frame. Afterwards, the players with small occlusions were tracked in goal scenes via template matching where the similarity measure should be high enough. The merge-split was also used for highly occluded blobs; however, error in locating occluded player was inevitable. The methods by Khatoonabadi and Rahmati (2009) and Intille and Bobick (1995b) were scarce works that applied the camera motion compensation. However, small number of features in broadcast soccer games makes it difficult to compute the image to model transformation. In Matsui et al. (1998), template matching was applied for tracking missed players after linking nearest extracted player blobs in successive frames. Although player templates could be updated with new matched templates (Yoon et al., 2002; Seo et al., 1997), imprecise or partially occluded player template might deviate from the real player template and lead to mistracking in subsequent frames. Accordingly, updating templates in occlusion situations should be prevented (Khatoonabadi and Rahmati, 2009). Moreover, the change in player's posture might decrease the estimated position precision. Also, template matching should be incorporated with motion or other constraints to deal with tracking nearby or occluding teammates in similar uniforms.

### 6.5. Meanshift and camshift

Meanshift is a non-parametric technique to find the mode of probability distributions, which first appeared in Fukunaga and Hostetler (1975), and it was later applied to player tracking. Accordingly, location and size of the search window and target model were initialized. Starting from the initial position, its neighborhood was iteratively searched for the best candidate to maximize the similarity measure (i.e., the Bhattacharyya coefficient between target and candidate color histograms). A few algorithms, such as ensemble tracking (Avidan, 2007) and camshift (Bradski, 1998), were expanded on this idea. Since the algorithm used a fixed size tracking window (Qian et al., 2007), target scaling could be performed by calculating the Bhattacharyya coefficient for different sizes and choosing the size with the highest similarity (Nummiaro et al., 2003). Another solution for size adaptation was camshaft, which was primarily introduced for head and face tracking (Bradski, 1998). An improved meanshift with motion prediction and player window enlargement was presented by Chiang et al. (2009). Each frame was converted into a probability distribution image from backprojection of the player histogram, and player tracking was performed via the improved meanshift. Furthermore, new players and outgoing players were detected by background subtraction near the image boundary and the low distribution area of the player window, respectively. As a result, the improved meanshift outperformed camshift. Since player maneuvers were difficult to represent with a single maneuver model, meanshift was employed for three motion models by Zhong et al. (2006). Then, one pseudo measurement from fusing the obtained measurements was used to drive IMM. Their method was more adaptive to agile motion of the players than only meanshift or combination of Kalman and meanshift. An obvious advantage of the meanshift and camshift over the template matching was avoiding brute force search. However, they could not handle total occlusions or occlusion among teammates without additional motion constraint. They also required that a portion of the player be inside the initial search window, and they might fail in the case of appearance changes (e.g. due to lighting changes) and players with similar colors of the background.

### 6.6. Kalman filter

In contrast to meanshift and template matching, motion model along with the player observation was inherently applied by KF (Barceló et al., 2005; Svensson, 2010). However, both the motion and measurement models were linear with additive Gaussian noise. Accordingly, any minor violation of the assumed motion model could be encoded in the process covariance matrix, which was hard to determine in the case of abrupt player and camera movements. Similarly, measurement errors could be considered in measurement covariance matrix, which assigned high values in the case of imprecise observations (e.g. occluding blobs). Moreover, the player state was usually defined by position, size and velocity. In Najafzadeh et al. (2015), one main player was tracked, and other players were tracked with respect to the main player by defining their relative position to the main player in the state vector of KF. The fusion of multiple features could also make the tracker more robust to occlusions and deformations (Misu et al., 2004, 2002). The method by Misu et al. (2002) consisted of several observations executed step by step. Player localization by color statistics, template matching, marking player head, square templates around characteristic points and size of the player bounding box were used in different steps. The algorithm automatically invalidated unstable results by increasing observation variances in KF covariance matrix. Moreover, tracking failures in highly congested area could be detected by monitoring the state covariance. The method by Xu et al. (2004a, b) and Ren et al. (2009) consisted of single-view and multi-view processing steps. The single-view step included player detection, labeling and tracking by KF in image space, which took advantage of partial observations in the update step. Finally, the KF was used by multi-view process to associate measurements from the single-view step. In Herrmann et al. (2014), player tracking was performed by finding local maxima on a confidence map by a greedy heuristic method using the KF prediction as the starting position. For this purpose, a single KF was applied for each tracked player. However, total occlusions were ignored. In Schlipsing et al. (2014), human intervention was applied for improving KF performance in challenging scenes. As linear system models were the base of conventional KF, erratically moving players and camera in soccer broadcast videos were problematic. The method by Xu et al. (2004a,b), Ren et al. (2009) and Misu et al. (2009) tracked players via KF through sequences captured by static cameras which alleviated at least the camera motion effects.

### 6.7. Particle filter

Regarding the limitation of KF in the case of nonlinear movements, PF (Arulampalam et al., 2002) also known as condensation or sequential Monte Carlo (SMC) has been widely applied. Particles representing possible solutions to the tracking problem (i.e., player candidates) were initialized, and their next states were predicted for propagating them. Then, each particle was weighted and the mean state of the player was estimated. Moreover, resampling was applied to avoid degeneracy problem. Output was also the particle with the largest weight or weighted average of all the particles (Vos and Brink, 2009). PF has been widely applied for player tracking (Hamid et al., 2010). A support vector regression (SVR) based reweighting scheme was applied by Zhu et al. (2006, 2005, 2007) to re-approximate the posterior density and avoid degeneracy in PF. As a result, the sample distribution for SVR particle filter was maintained much better than sampling importance resampling (SIR) particle filter. In order to handle occlusions, PSVC was combined with SVR by Zhu et al. (2009), which outperformed their previous work. Most of the related works made some assumptions about the player shape, such as a bounding box or an ellipse (Nummiaro et al., 2003), but the method by Dearden et al. (2006) represented the player by the collection of particles to reduce the size of the state space and the number of required particles. In order to reduce the number of particles, Rao-Blackwellization as a combination of KF and PF was presented by Hoyningen-Huene and Beetz (2009) in a data association framework in which particles were represented as configurations of all players. Moreover, the likelihood function played an important role in PF, since it determined the weights of particles. One popular likelihood function was color-based model where the color histogram was frequently employed with the Bhattacharyya coefficient as a similarity measure (Zhang et al., 2008; Nummiaro et al., 2003).

In order to deal with computation cost in a multi-feature space, hierarchical PF was presented by Vos and Brink (2009), Yang et al. (2005) and Wang et al. (2008) in which the computation was focused on more promising regions in a coarse-to-fine manner (cascade). In Vos and Brink (2009), Haar features and histogram of oriented gradients were applied for weighting particles in the first and second stage of the cascade, and the final weights were computed by multiplying particle weights from all stages. In contrast to Vos and Brink (2009), color was also used in Yang et al. (2005) to improve performance. In Wang et al. (2008), color, edge, and position information were applied. In contrast to most previous works which used discrete player blobs as entries of the tracker (Dearden et al., 2006), the method by Yao et al. (2010) applied a continuous vote-based confidence map in which each particle was weighted based on the map, HSV color histogram and local binary patterns. Moreover, estimated velocity based on optical flow with camera compensation was used at the prediction step instead of constant velocity models.

One of the main problems of PF lied in disability in tracking multiple players, since all the particles might migrate to one of the modes. As a solution, mixture of particle filters (PFs) was applied by Vermaak et al. (2003). In Wu et al. (2008), BIDPF was applied in which the boosting proposal detected incoming players, and the interactively distributed particle filter (IDPF) handled occlusions. The method by Needham and Boyle (2001) incorporated KF with PF to avoid a player samples from splitting when players moved closely. In Du and Piater (2007), a player was tracked from multiple cameras in both ground and image planes by PFs which interacted with each other to refine the results. In Chai et al. (2011), multiple independent PFs were applied while the filter weight decreased during occlusions, and the samples propagated without resampling. In contrast to most PF-based trackers which assigned individual PF to each player, tracking was performed by propagating samplesets (group of players) within the supersampleset (a collection of sample sets), and the sampleset with the highest score was selected. Afterwards, one KF was assigned to each player to update the observed values in the best sampleset. Another approach called a dual-mode two-way Bayesian inference was presented by Xing et al. (2011), which switched between isolated and multiple occluded players tracking recognized by undirected graph. The isolated player was tracked with a general observation model, and the occluded players were tracked with a dedicated observation model for each player. Moreover, missed observations were handled by backward smoothing. Their framework outperformed standard PF and two-stage tracking algorithm. In Baysal and Duygulu (2016), instead of assigning separate particles to each player, densely sampled particles on the field model were shared among all the players. Then, interactions among the players were considered by globally evaluating likelihood of locating players on the particles using combined appearance and motion model and a color-based occlusion detector. It also outperformed meanshift, optical flow, color-based PF and color-based mixture PF, K-shortest path (KSP), dynamic program tracker (DP) and TMCNF. The method by Nummiaro et al. (2003) compared three methods, namely, color-based PF, meanshift and Kalman-meanshift. Although Kalman-meanshift reduced meanshift iterations, it failed during nonlinear movements. Contrastingly, nonlinear movement was dealt by PF, and tracking was more reliable due to its multiple hypotheses. However, a more precise localization was achieved by the other two trackers, since the location was estimated by the mean value of particles in PF. In Davis (2008), KF outperformed nearest neighbor. While noisy measurements were not compensated by nearest neighbor and nonlinear player movements were problematic for KF, PF compensated for these limitations. Moreover, multiple trackers were avoided to associate with the same player by assigning low weights to the states occluded with other trackers.

The main drawback of PF was its dependence on the number of particles. The large set increased computational cost, and the small set might result in non-accurate player tracking (especially player localization). Moreover, it was unable to deal with total occlusions and nearby teammates, since it could not restrain the particles of a player from attracting to the other player. As reviewed earlier, there have been a bunch of papers to present solutions for alleviating these shortcomings.

### 6.8. *Temporal spatio-velocity (TSV) transform*

TSV was introduced by Sato and Aggarwal (2004) for human tracking, and it was later applied for player tracking (Sato and Aggarwal, 2005). In order to perform TSV transform, spatio-temporal image was converted into a spatio-velocity plane (TSV image) by Hough transform. Therefore, pixel velocity could be found by finding local maxima in the TSV image, and the blobs could be generated by thresholding the TSV image and grouping pixels with similar velocities. In order to track players, two tracking methods were applied, namely, simple blob tracking based on nearest distance for isolated players and local TSV for occluding players. Simple operations, noise suppressive and separating occluding blobs with different velocities were the main advantages of the TSV. However, low image resolution in soccer sequences may be a stumbling block.

### 6.9. *Optimizing an objective function*

Player tracking can also be performed by optimizing an objective function. The method by Mackowiak (2013) defined the cost function based on the size and overlap area between bounding boxes of tracked player and detected ones. Then, the candidate box with minimal cost was selected as the same player. Moreover, the prediction was applied whenever a tracked box could not be matched. Evaluation of the maximum a posteriori probability (MAP) (Mazzeo et al., 2008) and maximization of a joint probability model reflecting players' motion and appearance (Kang et al., 2004) were also applied for player tracking. In Khan and Shah (2009), the foreground likelihood maps obtained by background subtraction from all views were fused to represent the likelihood of players occupying the locations on the scene. Then, occupancies were used to define and minimize energy function for tracking players. Although track identities might be switched due to the lack of appearance information, the method outperformed appearance-based approaches for players in similar color uniforms. Following playfield and player detection, a linear prediction model regarding SSM based on the distance between blobs was applied by Duh et al. (2013). In Martín and Martínez (2013), the distance between blob centroids and their colors in a match matrix were applied for corresponding the blobs in subsequent frame. In D'Orazio et al. (2009), a tracking-by-detection approach was proposed in which a similarity measure based on the blob position, velocity, dimension and appearance was defined. Therefore, the maximum confidence gave the best multi-player configuration model. It was also possible to judge single player, occluding blob, new entry player, outgoing player, non-segmented blob and resumed blob after disappearance by predicting positions.

## 6.10. Data association methods

Multi-player tracking involves the problem of data association to determine which measurements are generated by which players. Accordingly, a validation gate can be placed around the player to reduce association candidates. Several data association algorithms have been proposed. They can be divided into single frame association methods which make assignments based only on the current frame (e.g. nearest neighbor (NN), global nearest neighbor (GNN)) or multi-frame association methods (e.g. MHT). They can also be divided into single target association (e.g. NN) or multi-target association (e.g. GNN, JPDA and MHT). Moreover, the choice of the association algorithm depends on the particular application. Most player tracking algorithms ignored association or assumed that the association was trivial so that NN could be effective. Different association approaches that addressed soccer player tracking are reviewed in the following subsections.

### 6.10.1. Nearest neighbor association

Two most common single frame association methods were NN and GNN (Svensson, 2010). In NN, each measurement was assumed to originate from the closest predicted player, and it can be defined using the Mahalanobis distance (Cox, 1993). The method by Barceló et al. (2005) applied NN for data association along with KF tracker. In Junior and Anido (2004), the nearest player to the player position in the previous frame was used for association, and color information was applied for multiple nearest players. An obvious advantage of NN was its simplicity; however, it might associate several closely spaced players with the same measurement. GNN (Zhu et al., 2009) was also a global version of the NN by considering simultaneous assignment of all players under the constraint that an observation could be associated with at most one track. However, it still might fail in the case of closely spaced players and high number of false measurements.

### 6.10.2. Probabilistic data association (PDA) and joint probabilistic data association (JPDA)

One approach proposed to improve GNN performance was JPDA (Fortmann et al., 1983) in which a track was updated by a weighted sum of all observations in its gate. In other words, an observation could contribute to the update of more than one target. The PDA dealt with multiple targets independently; however, measurements were evaluated jointly by JPDA as an extension of the PDA. The main limitation of the original JPDA was its inability to perform track initiation and deletion. Therefore, it was appropriate when the number of tracks was known and remained fixed throughout the sequence which was not the case in soccer broadcast videos. The method by Abbott and Williams (2007) applied JPDA with KF for tracking players in world coordinate system in sequences captured by static cameras.

### 6.10.3. Multiple hypothesis tracking

MHT was originally developed by Reid (1979) for tracking multiple targets in clutter. Its version by Cox and Hingorani (1996) was later applied in soccer sequences (Beetz et al., 2006, 2007; Gedikli et al., 2007). MHT is a statistical association algorithm with capabilities of track initiation, track termination, track continuation, spurious measurement handling and uniqueness constraint preservation. The data association is also postponed to the later time step for resolving uncertainties. Generally, MHT includes hypothesis matrix creation, hypothesis generation, calculation of hypothesis probabilities and KF associated with each target and hypothesis management (Cox and Leonard, 1991; Blackman, 2004). The number of look forward frames (e.g. 15 frames) depended on the distinctiveness of the observations in Gedikli et al. (2007). Moreover, making assumptions about the specific application of soccer sequences (e.g. new tracks can be created at the borders of the frame) could enable pruning the hypotheses tree (Beetz et al., 2006). An improvement of MHT using a modification of Murty's algorithm was presented in Joo and Chellappa (2007) by permitting association of one player with multiple measurements and vice versa. As a result, the multiple hypotheses outperformed the single hypothesis in the case of occlusions and noise. However, the tracking error occurred due to sudden change of the player velocity in severe occlusions or a bad estimate of measurements within an occluding blob. The method by Beetz et al. (2007) also refined MHT to assign more than one player to a measurement. The MHT yielded better results than methods with one association hypotheses; however, large computation cost depending on the level of ambiguity in application can be problematic for real-time applications.

### 6.10.4. Markov chain Monte Carlo data association

Although most data association algorithms (e.g. JPDA and MHT) assumed one to one correspondence between observations and players (which can be violated in the case of occlusions or false alarms), MCMCDA (Oh et al., 2004) did not consider such assumption. Generally, MCMC (including algorithms such as Gibbs sampling and the Metropolis-Hastings) approximated solution of a combinatorial optimization problem by randomly searching the space instead of enumerating all possible associations. Moreover, some types of moves (birth/death, extension/reduction and split/merge, segmentation and aggregation) were utilized by the proposal distribution of MCMCDA to compute the state change probability (Liu et al., 2009; Zhang et al., 2008; Tong et al., 2011). Multi-player tracking with MCMCDA was presented by Liu et al. (2009) and Tong et al. (2011) in which the whole detection and labeling results were taken as observations. Then, the best association was represented by configuration on a neighborhood graph which optimized Gibbs distribution. For this purpose, Metropolis-Hastings was adopted to estimate the optimal solution, and observations of missed players were interpolated. In Zhang et al. (2008), the extracted player trajectories by PF were refined offline by MCMCDA incorporated with Metropolis-Hastings sampling, since occlusions and false alarms made one trajectory not corresponded to one player. MCMCDA was also capable of initiating and terminating a varying number of tracks, and it was flexible to incorporate the specific knowledge of an application. Although MCMCDA outperformed MHT in terms of accuracy and efficiency under extreme conditions in Oh et al. (2004), there were several parameters to be set. Moreover, long occlusions, serious video blur, abrupt camera motion and player tangle might still lead to failure.

## 6.11. Other developments and combinatorial approaches

The association was applied implicitly in some papers, for instance, by decreasing the weight of each predicted particle overlapped with predicted adjacent player in PF (Ok et al., 2002). Moreover, there have been papers concerned with jointly modeling target state, target existence, data association and interactions among targets within a multi-target tracking framework in which maximum a posteriori (MAP) estimation of the joint target state could be approximated. Accordingly, SMCJPDA was presented by Zheng and Xue (2009) where data association was modeled within the SMC, and player detection and tracking were unified to track a varying number of players. A similar approach was adopted by Vermaak et al. (2005) in which target existence was treated jointly with the target state and association within a probabilistic framework. It was then solved by a mixture KF for tracking unknown and variable number of players.

The reviewed trackers were categorized regarding the main approach. However, different trackers were combined by some

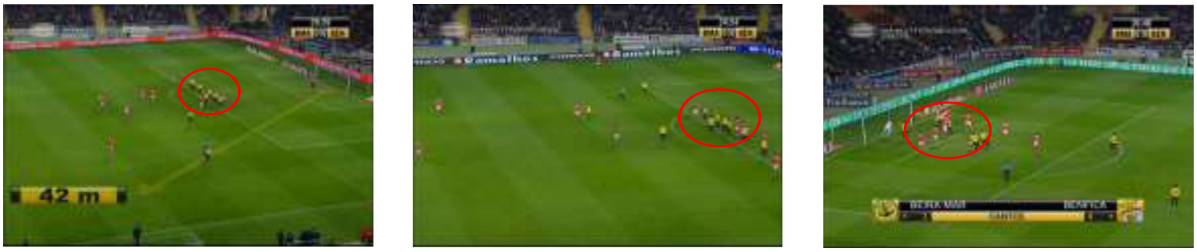


**Fig. 9.** Scenes with heavily occluded players.

authors (KF and point tracker (Mathes and Piater, 2006), PF and graph (Itoh et al., 2012), KF and template matching (Misu et al., 2002), Rao-Blackwellization as the combination of KF and PF (Hoyningen-Huene and Beetz, 2009), KF and PF (Needham and Boyle, 2001), KF and JPDA (Abbott and Williams, 2007)) to improve the tracking performance. The main works published in player tracking are summarized in Table 3. The applied method, real world tracking (on the ground plane), number of cameras (single or multi-camera), association method, camera status (moving or fixed camera) and the number of test frames are presented in the columns.

It is difficult to compare methods due to the variability in datasets particularly for broadcast sequences, lack of access to the earlier methods' codes, different evaluation criteria and lack of evaluation process in the literature. Accordingly, quantitative evaluation has been ignored by some previous works, or it has been performed via few visual demonstrations. In addition, the weaknesses of tracking algorithms were rarely discussed in their corresponding papers. Some strengths and weaknesses of tracking frameworks are summarized in Table 4. Real world tracking, considering data association and sufficient number of test frames are strengths, which were denoted in Table 3. Moreover, some strengths and weaknesses are missed in Table 4; since they were not clearly addressed in the literature. Although failure in occlusion among more than two players was denoted by few authors, it might occur in most trackers. Imprecise player localization was also deduced regarding the qualitative results. Moreover, the delay in tracking occurred due to working on a batch of frames or looking forward and backward frames (particularly in graph-based tracking) (DL-Batch). Although real world tracking is an advantage, it can be a weakness in broadcast videos due to lack of features in the middle of the field. As a result, the tracking results were significantly affected by errors in image to model registration. Moreover, tracking-by-detection methods (T-by-D) were directly affected by player detection or labeling which were solved independently.

## 7. Occlusion resolution

Occlusion occurs when some players are located in front of the others along the optical axis of the camera, and thus backward players are hided partially or completely. Consequently, information is mostly available for the visible players and occluded ones may not be identified. Occlusion is the most challenging problem in tracking soccer players. It is sometimes so severe that even the user cannot recognize the occluded player. The image quality is also very important to deal with this problem. Occlusion is very common in soccer sequences, and it may occur among teammates, competitors and referee. Since a defender is often close to the offensive player of the other team, most occlusions occur among competitors, and distinctive appearance of competitors can be used as the solution. However, extremely difficult situation arises due to the occlusion among multiple teammates with similar clothes or heavily grouped players during some free kicks or corners (Fig. 9). Since extremely sophisticated trackers would eventually lose the identity of a track due to the heavy occlusions, it is necessary to address occlusion during tracking multiple players. Although some authors have tried to track group without division, it is crucial to locate players in an occluding blob for high level analysis of the soccer games. Occlusion detection and different approaches to deal with occlusions are reviewed in the following subsections.

### 7.1. Occlusion detection

At the first step, it is essential to define some criteria indicating whether an occlusion is present or not. The drastic increase in the area of the player blob between the previous and current frame (Iwase and Saito, 2003), decrease in the number of labels around the tracking player (Iwase and Saito, 2002) and the dominant player scale learnt from the scale histogram of contours in the playfield (Zhang et al., 2008) have been applied for occlusion detection. One limitation of using size constraints for occlusion detection was size variability due to the player distance from camera and failure of detecting total occlusions. The specific structure of the graph could also help occlusion detection (Manafifard et al., 2015). Moreover, occlusion has been detected when the predicted players fell close (e.g. less than the width of the region) (D'Orazio et al., 2009; Sato and Aggarwal, 2005), the predicted positions were within the same foreground blob (Xu et al., 2004a, Utsumi et al., 2002) or the regions of none of the players could be extracted (Ohno et al., 2000). Trained SVM was also used by Choi and Seo (2011) to estimate the number of people inside a blob, and the feature vector for SVM consisted of the camera index, width, height, area and bottom position of a blob. In Baysal and Duygulu, 2016, occlusion was detected by classifying the color likelihood with a trained Bayes classifier on jersey samples. Moreover, point trackers detected the total occlusion by the number of matched points between two players in subsequent frames (Gabriel et al., 2005).

### 7.2. Occlusion handling

Short connections between nearby players could be eliminated by morphological operations (Figueroa et al., 2006; Khatoonabadi and Rahmati, 2009); however, this solution might lead to eliminating player feet or player localization errors. Some attempts have been made at using multiple cameras (Xu et al., 2004a, Junior and Anido, 2004; Iwase and Saito, 2003; 2002; Li and Flierl, 2012). However, this kind of solution was expensive while occlusions were not totally solved (Figueroa et al., 2006). The method by Misu et al. (2009) used two types of cameras one for player tracking and the other for following occluding players. Then, occlusion was resolved using face and back-number recognition. Although pixel intensities were used by few authors to track a player during occlusion (Intille and Bobick, 1995b), color cues were more reliable. In Ohno et al. (2000, 1999), an occluding player was determined using color, and the occluding player neighborhood was searched for the occluded player. Moreover, the lower player (closer to the camera) in an occluding blob corresponding to teammates was

**Table 3**
A review of player tracking (RW: real world tracking, DA: data association method, S/M: single/multiple cameras, M/F: moving (with pan or tilt or zoom) /fixed camera, Y: yes, N: no, NN: nearest neighbor, SR: soccer, IN: indoor sequence, seq: sequences, GT: ground truth, min: minutes, sec: seconds).

| Ref. | Tracking method | RW | S/M | DA | M/F | Frames |
|---|---|---|---|---|---|---|
| Martín and Martínez (2013) | Match matrix | Y | M | - | F | 2 min of the match |
| Duh et al. (2013) | SSM | N | S | - | M | 1600 |
| Herrmann et al. (2014) | Finding local maxima on a confidence map using the KF prediction as the starting position | N | S/M | - | M/F | 2 broadcast clips (925 + 752), 2 clips from static cameras (3000 + 2500) |
| Ben Shitrit et al. (2014) | MCNF and TMCNF on DAG and linear programming | Y | S/M (SR) | MCNF and TMCNF on DAG | M/F (SR) | SR (3000 × 6), basketball (1500 + 4000 + 5500) and PETS'09 dataset (795) |
| Liu et al. (2009) | MCMC | N | S | MCMC | M | First clip (100), Two other clips (250) |
| Khatoonabadi and Rahmati (2009) | Template matching and merge-split | Y | S | - | M | Seq that were totally 8 min and 1/4 of frames (15,411 players) was evaluated |
| Misu et al. (2009) | KF and Bayesain formulation | Y | M | NN | Both | - |
| Xu et al. (2004a) and Ren et al. (2009) | KF | Y | M | NN | F | 2 seq of 5000 frames |
| Zheng and Xue (2009) | SMCJPDA | N | S | SMCJPDA | M | SR (3234), pedestrians (2000) |
| Chiang et al. (2009) | Improved meanshift | N | S | - | M | 5 SR videos: Australia versus Italy (190) |
| Zhong et al. (2006) | IMM combined with meanshift | N | S | - | F | - |
| Joo and Chellappa (2007) | MHT | N | M | MHT | F | Effectiveness validation of SR (600), quantitative evaluation of SR (2500), two non-SR seq (180) |
| Khan and Shah (2009) | Minimizing occupancy-based energy function | Y | M | - | F | Parking lot (over 3000), IN (-), basketball (1000), SR (1000) |
| Abbott and Williams (2007) | JPDA-KF | Y | M | JPDA | F | 100 |
| Barceló et al. (2005) | KF and graph | N | S | NN and graph | M | - |
| Xing et al. (2011) | Dual-mode two-way Bayesian inference | N | S | - | M | Tracking occluding blobs (40), 5 football (2593), 5 basketball (1572) and hockey (at least 100 frames) seq |
| Mackowiak (2013) | Optimizing cost function and motion vector | N | S | - | M | - |
| D'Orazio et al. (2009) | Tracking-by-detection using similarity measure | Y | M | - | F | One track of goalkeeper (2883), 45 tracks of players (15,226) and group blobs (eight seq of 3000 frames) |
| Dearden et al. (2006) | PF | N | S | - | M | At least 250 frames |
| Du and Piater (2007) | PF | Y | M | - | F | Surveillance and SR seq (-) |
| Baysal and Duygulu (2016) | Improved PF (Sentioscope) | Y | M | Implicitly | F | 900 min without GT, 150 sec with GT (750), ISSIA (3000) |
| Wu et al. (2008) | BIDPF | N | S | - | M | 2 video clips (Second clip contained 1047 frames) |
| Vandenbroucke et al. (1997a) | Snake | N | S | - | F | 2 SR games (5) |
| Lefèvre et al. (2000) | Snake | N | S | - | M | 100 |
| Lefèvre and Vincent (2004) | Snake | N | S | - | M | - |
| Wang et al. (2006) | PF based SMOG, PF and meanshift based on color histogram | N | S | - | - | - |
| Nummiaro et al. (2003) | Color-based PF, Meanshift, Kalman-meanshift | N | S | - | M | Basketball (-), snowboarder (more than 60), mock surveillance (450), tracking one SR player (over 438), moving stairs (-), traffic (234), face (600) |
| Pallavi et al. (2008) | Optimal path search using dynamic programming in graph | N | S | Graph | M | 1193 |
| Figueroa et al. (2006) | Minimal path searching in graph | Y | M | Graph | F | 10 min |
| Sullivan and Carlsson (2006) | Graph | Y | M | Graph | F | About 1000 |
| Manafifard et al. (2015) | Graph and PSO | Y | S | Graph | M | 491 |
| Mentzelopoulos et al. (2012) | Motion detection by checking pixel distribution based on entropy on different directions | N | S | - | F | 17999 (SR), 95058 (non-SR) |
| Vos and Brink (2009) | Hierarchical PF | N | S | - | F | 100 |
| Yang et al. (2005) | Hierarchical PF | N | S | - | M | SR (at least 200), non-SR (-) |
| Wang et al. (2008) | Hierarchical PF | N | S | - | M | 5 seq: one seq included 73 frames |
| Zhu et al. (2006) | Support vector regression PF | N | S | - | M | 3599 |
| Zhu et al. (2009) | Combined PSVC with support vector regression PF | - | S | GNN | M | - |
| Beetz et al. (2006) | MHT | Y | S | MHT | M | - |
| Gabriel et al. (2005) | Interest point tracking | N | S | - | SR (M), IN (F) | SR (50), IN (200) |

(continued on next page)

**Table 3** (*continued*)

| Ref. | Tracking method | RW | S/M | DA | M/F | Frames |
|---|---|---|---|---|---|---|
| Hayet et al. (2005) | Interest point tracking by point distribution models (2D tracking), KF and association (3D tracking) | Y | M | Combinatorial optimization | M | 400 |
| Mathes and Piater (2006) | KF and point distribution manifolds | N | S | - | M | SR (150), PETS 2001 surveillance video (320) |
| Li and Flierl (2012) | Matching SIFT features | Y | M | - | F | 1500 |
| Intille and Bobick (1995a) | Template matching and closed world method | Y | S | - | M | 270 |
| Seo et al. (1997) | KF and template matching | Y | S | - | M | 150 |
| Yoon et al. (2002) | Template matching | Y | S | - | M | 40 |
| Sato and Aggarwal (2005) | Local TSV (occluded blobs) and simple blob tracking (isolated blobs) | Y | S | Local TSV (-), simple blob tracking (NN) | M | 870 |
| Vermaak et al. (2005) | Mixture Kalman to solve joint probabilistic data association | N | S | JPDA | M | - |
| Hamid et al. (2010) | PF (in each view) and K-partite graphs (fusing multiple views) | Y | M | - | F | 60,000 |
| Iwase and Saito (2003) | NN based on distance, label, area and fusing information of cameras by averaging using homography | Y | M | NN | F | scene 1 (450), scene 2 (150) |
| Iwase and Saito (2002) | NN (2D tracking) and fusing information of cameras using fundamental matrix (3D tracking) | Y | M | NN | F | 2 scenes (350 + 190) |
| Needham and Boyle (2001) | PF and KF | Y | S | Propagation of samplesets | F | Every 5 frame from 835 frames |
| Misu et al. (2004) | KF | Y | M | - | F | 7 sec seq (about 210) |
| Zhang et al. (2008) | PF incorporated by MCMC | N | S | MCMC | M | 1050 |
| Ok et al. (2002) | PF | N | S | Implicitly | M | - |
| Choi et al. (2004) | PF | N | S | - | M | - |
| Chai et al. (2011) | Multiple independent PFs | N | S | - | M | - |
| Hoyningen-Huene and Beetz (2009) | Rao-Blackwellized Resampling PF (RBRPF) | N | M | Sampling associations | M | 6988 |
| Davis (2008) | NN, KF, PF | N | S | Improved tracker | M | 200 |

considered to be an occluding player. Occlusion among more than two players was also handled using the average velocity, positional relationship and color information. The method by D'Orazio et al. (2009) searched the occluding blob from the predicted position for the player color. However, performance decreased when the number of occluding players became more than three. Color classification of pixels in a hybrid color space (Vandenbroucke et al., 1997a), color histogram backprojection (Seo et al., 1997) and template matching using isolated players' templates from the nearest frames (Bebie and Bieri, 1998) have also been applied for resolving occlusions among competitors. In Khatoonabadi and Rahmati (2009), histogram backprojection, template matching and split-merge were applied for occlusions between players and television logos or advertisement boards, small occlusions and large occlusions, respectively. The main limitation of the color-based methods was their disability to resolve occlusions among teammates. Furthermore, they attempted to track the distinguishable parts of the player, and the positioning of the occluded players was imprecise.

The main clue for resolving occlusion among teammates or nearly total occlusion was prediction using a motion model which usually reflected the constant velocity assumption (Chiang et al., 2009; Khatoonabadi and Rahmati, 2009). However, it was vulnerable to any violation of the assumption. The method by Joo and Chellappa (2007) searched for players in an occluding blob using the prediction and the constraint that each side of the observation bounding box should be tangent to at least one player. There were also trackers which handled occlusions using predictions along with observations (e.g. KF and PF). Since PF could not restrain a player from attracting particles of nearby teammates, the likelihood of each particle was weighted by occlusion alarm probability in Ok et al. (2002). Therefore, the particle weight became low when it was closer to the predicted adjacent player than the predicted original one. An improved meanshift using motion prediction was also proposed by Chiang et al. (2009). The occluding player was the peak of the probability distribution of the search window, and the occluded player was searched around the occluding one in the case of occlusion among competitors. In Xu et al. (2004a, b) and Ren et al. (2009), partial observations (e.g. due to occlusion) were used for updating the player state by KF, which outperformed those without any partial observations. For more occluding players, the blob was forwarded to the multi-view processing stage due to the uncertainty. Backward smoothing process (Xing et al., 2011) and backtracking scheme (Pallavi et al., 2008) were also applied for mistracking due to occlusions or abrupt player movements. Moreover, the occlusion resolution was performed by temporal information using TSV (Sato and Aggarwal, 2005) or graph representation (Figueroa et al., 2006; Sullivan and Carlsson, 2006; Barceló et al., 2005; Manafifard et al., 2015). The method by Barceló used graph representation and split-merge table based on KF for handling occlusions, as all the joins and splits could be represented in the graph. The split-merge method was also applied by Duh et al. (2013) and Najafzadeh et al. (2015). Moreover, an improved PF was proposed by Baysal and Duygulu (2016) to handle occlusions, which associated particles with the occluded tracks regardless of their weights. Generally, few works have focused on resolving occlusion among more than two players which is still an open problem. The main approaches presented for occlusion resolution are summarized in Table 5.

## 8. Evaluation

Since appropriate evaluation is required for comparison of algorithms, different criteria have been applied by the previous works on different datasets for qualitative and quantitative eval-

**Table 4**
A review of strengths and weaknesses of tracking frameworks (F-IN: fusing information across multiple cameras, HC: using hard constraints, OC: occlusion, H–OC: handling occlusion, L-OC: long term occlusion, DL-Batch: delay in tracking, RE-C: reducing computational cost, NOT-IS-OC: not isolating partially occluding players in an occluding blob, T-by-D: tracking-by-detection method, E-CM: eliminating camera motions effect, COMB-TR: improving performance by combining the trackers, IM-LOC: imprecise player localization, NL-MOV: nonlinear and abrupt player or camera movements, H-Match: failure in detecting or matching interest points for distant or blurred players, TR: tracking players, AP: appearance, F-B: look forward or backward, D-N-P: depending on the number of particles, IG: ignoring, Outp: outperformed, AF: tracking results were affected by, -: unclear or unmentioned).

| Ref. | Strengths | Weaknesses |
|---|---|---|
| Martín and Martínez (2013) | Simple, F-IN | HC, semi-supervised |
| Duh et al. (2013) | Simple | HC and IG OC |
| Herrmann et al. (2014) | Fast, simple, using both gradient and color features | Using a greedy heuristic method, IG DA and total OC |
| Ben Shitrit et al. (2014) | Fast (MCNF (3.95f/s), T-MCNF (187.5f/s)), well assessed by GMOTA, Outp KSP, C-KSP and DP, commercialized in different sports | DL-Batch |
| Liu et al. (2009) | Considering track length in several frames, player label and motion consistency for TR | T-by-D, large number of parameters, imprecise H–OC by interpolation, failure in L-OC and NL-MOV, DL-Batch |
| Khatoonabadi and Rahmati (2009) | E-CM | AF by registration, NOT-IS-OC, limited to the goal scenes |
| Misu et al. (2009) | Integrating player trajectories and identities from a fixed wide angle camera and a motion-controlled camera for H–OC | Requiring a special apparatus for measuring and controlling camera parameters |
| Xu et al. (2004a) and Ren et al. (2009) | Real-time, F-IN, good coverage of the pitch using multiple cameras, which helped H–OC | IG NL-MOV in KF, AF by inaccurate calibration due to the low textured pitch |
| Zheng and Xue (2009) | Integrating the detection and tracking information, working well in tracking a varying number of targets, H–OC even in the case of low detection rate | - |
| Chiang et al. (2009) | Improving meanshift for size adaptation and H–OC | It required that a portion of the player be inside the initial search window |
| Zhong et al. (2006) | Using multiple models, more adaptive to NL-MOV than meanshift and combination of KF and meanshift | It required studying parameter selection, tracking a single player |
| Joo and Chellappa (2007) | Improving MHT for H–OC and fragmented players, real time, evaluation regarding OC | IG AP cues |
| Khan and Shah (2009) | TR in crowded and cluttered scenes, H–OC using an image-based method and multiple views | Increasing computation cost with increase in number of views, planes and image resolution, the highest used image resolution was 576 × 720 due to the memory issues, IG AP cues, switching identities in the case of heavy and L-OC |
| Abbott and Williams (2007) | RE-C using LBSCCA | Failure when new tracks appeared at unpredictable locations, unsuitable for moving camera |
| Barceló et al. (2005) | TR on image mosaic (E-CM) and matching the mosaic with the field model, COMB-TR | DL-Batch, AF by errors in the mosaic construction, not robust H-OC |
| Xing et al. (2011) | Developing a progressive observation model, F-B for H–OC, TR in different sports, Outp PF and two-stage tracking algorithm, handling abrupt pose changes of the player, fast (15 fps) | Requiring the training, DL-Batch, missing one of the occluding players in serious OC |
| Mackowiak (2013) | Simple | T-by-D, HC, improper H–OC |
| D'Orazio et al. (2009) | Real-time, re-identify the player after a merge | It occasionally failed to correctly detect players within the occluding blob |
| Dearden et al. (2006) | Fast (5fps), H–OC, no prior assumption about the shape or size of the player | Sharing no information between PF trackers might cause tracking failure for occluding teammates, D-N-P |
| Du and Piater (2007) | Collaborating PFs in multiple cameras and ground plane, which relaxed the dependence on precise foot positions | D-N-P |
| Baysal and Duygulu (2016) | H-OC, global evaluation of the likelihood, real-time (65ms ± 6 ms per frame), TR in different matches. Outp meanshift, optical flow, color-based PF, color-based mixture PF, KSP, DP and TMCNF | D-N-P |
| Vandenbroucke et al. (1997a) | Tracking the player contour, using color classification with snakes for H–OC | Color classification of players with similar colors in their uniforms would be unreliable |
| Lefèvre et al. (2000) | Tracking the player contour | IG H–OC among two players or among player and line |
| Lefèvre and Vincent (2004) | Using a splitting process to deal with OC, computing gradient only on the area of interest, using a multiresolution framework to RE-C | Sensitive to pixels with high gradient values in background, disappearance of small players in low resolutions |
| Wang et al. (2006) | Improving AP model, Outp meanshift and condensation based on color histogram | D-N-P |
| Nummiaro et al. (2003) | H-OC, real-time, tested on different scenarios, comparing 3 trackers (Kalman-meanshift reduced meanshift iterations, NL-MOV was dealt by PF) | Kalman-meanshift failure during NL-MOV, IM-LOC in PF, tracking a single player, gaps during L-OC |
| Pallavi et al. (2008) | F-B for generating modified tracks of occluding players, robust for small or occluded object in a noisy environment, tested on other applications (e.g. tracking infant body parts, gait biometry) | DL-Batch, IM-LOC, increasing computation with similar objects in the environment, NOT-IS-OC |
| Figueroa et al. (2006) | H-OC using a graph which allows the F-B, computing variables (e.g. covered distance, velocity) using the TR | Increasing computations by increase in number of F-B frames in L-OC, not H–OC among more than 3 players, IG color, DL-Batch |
| Sullivan and Carlsson (2006) | H-OC using continuity of motion, AP and depth ordering of the players (more precise positions for players in occluding blobs could be achieved) | Offline |
| Manafifard et al. (2015) | RE-C in the case of TR regarding multiple frames by partial exploration of the search space, isolating partially occluding players | T-by-D, AF by registration, HC, DL-Batch |

(continued on next page)

**Table 4** (*continued*)

| Ref. | Strengths | Weaknesses |
|---|---|---|
| Mentzelopoulos et al. (2012) | No need for camera calibration or background extraction, fast | IM-LOC |
| Vos and Brink (2009) | RE-C by a hierarchical approach using fast descriptors and several features | Using a grayscale video, not H–OC, missing still or slowly moving player for a long time, D-N-P |
| Yang et al. (2005) | Real-time, RE-C by a hierarchical approach using several features | IM-LOC, D-N-P |
| Wang et al. (2008) | RE-C by a hierarchical approach using several features, H–OC, Outp meanshift and meanshift-PF | IM-LOC, D-N-P |
| Zhu et al. (2006) | Improving PF for a smaller sample set | - |
| Zhu et al. (2009) | Outp Vos and Brink (2009), OC was solved more effectively compared to Vos and Brink (2009) | |
| Beetz et al. (2006) | H-OC in MHT | IG color cue in MHT |
| Gabriel et al. (2005) | Robust to OC | The linear movement assumption for a lost player might be violated, IG objects entering or leaving the scene, H-Match |
| Hayet et al. (2005) | Robust to OC | H-Match, AF by errors in real world tracking |
| Mathes and Piater (2006) | Robust to partial OC, learning the player model incrementally, considering the local AP and the spatial configurations of points, TR through scale, AP and shape changes, as long as they exhibit sufficient texture | H-Match |
| Li and Flierl (2012) | H-OC, improving the reliability of TR compared to SIFT-based 2D tracking using 3D information, real-time | H-Match |
| Intille and Bobick (1995a) | E-CM | HC, large processing, IG color cue, failure when the player was close in AP to nearby object, improper H–OC, AF by registration |
| Seo et al. (1997) | COMB-TR | IG OC among teammates |
| Yoon et al. (2002) | Tracking on the mosaic and field model | HC, IG OC among teammates, NOT-IS-OC |
| Sato and Aggarwal (2005) | Simple operations, noise suppressive, isolating occluding blobs | The low resolution of an image could be a stumbling block |
| Hamid et al. (2010) | F-IN | D-N-P in PF |
| Iwase and Saito (2003) | F-IN for H–OC | Failure in OC among more than two players, errors in F-IN, IG motion information, limited to the penalty area, HC |
| Iwase and Saito (2002) | F-IN to handle OC | Failure might occur when the player was occluded by two or three players, AF by errors in fundamental matrix, HC in NN |
| Needham and Boyle (2001) | COMB-TR, metric evaluation | AF by registration result, D-N-P in PF |
| Misu et al. (2004) | Using multiple features | IG NL-MOV in KF |
| Zhang et al. (2008) | COMB-TR | Offline trajectory refinement, D-N-P in PF |
| Ok et al. (2002) | H-OC among teammates in PF | Missing occluded player in an initial occluding blob after OC, D–N-P |
| Chai et al. (2011) | Real-time | IM-LOC, using independent PFs |
| Davis (2008) | Comparing 3 methods (PF Outp KF and NN, and KF Outp NN), PF was suitable for NL-MOV | AF by errors in player detection, D-N-P in PF, IG NL-MOV in KF, HC in NN |

uation of the performance. Most of them have tried to depict the performance subjectively in which results (e.g., bounding boxes and identities) were displayed on the original test sequences. It was also necessary to establish a ground truth for quantitative evaluation. The ground truth has been commonly computed by manually labeling player positions (e.g. superimposing bounding boxes on players), corresponding classes and trajectories using a graphical user interface. Accordingly, an interpolation could be used to automatically assign the players' positions between consecutive manual assessments. The applied datasets and evaluation criteria for described steps of player tracking are reviewed in the following subsections.

### 8.1. Datasets

Soccer broadcast videos including different types of games (World cup, Olympic games and League matches) were often recorded from broadcast television programs (Zhu et al., 2006; Zhu et al., 2009; Manafifard et al., 2016; Hoernig et al., 2015; Sun and Liu, 2009; Gerke et al., 2013; Zhu et al., 2007), and static cameras were usually set up by authors (Misu et al., 2009). Moreover, some previous datasets (Choi and Seo, 2011; Khan and Shah, 2009) are not publicly available anymore. As a result, experiments were conducted using few available datasets from static cameras (VS-PETS 2003 dataset (University of Reading 2016), ISSIA (Spagnolo) dataset (D' Orazio et al., 2009)). VS-PETS 2003 dataset consists of three synchronized views (each clip is 2499 frames long) with the image size of 720 × 576. The sequences were captured from the cameras positioned in different corners of the pitch. Conversely, the ISSIA consists of six synchronized views captured by six Full-HD cameras, and three cameras were placed on each major side of the pitch. Therefore, six clips from this dataset at 25 fps (each clip is 3002 frames long) with the image size of 1920 × 1088 cover the whole pitch. ISSIA provides the position of the players and referees in each frame, and the first 300 frames of each sequence have been labelled for initializing background subtraction algorithms. Some sample frames from each view of the VS-PETS 2003 and ISSIA are shown in Fig. 10. One sequence of the VS-PETS-2003 dataset captured from the third camera (view 1 in Fig. 10) was used by Joo and Chellappa (2007), Herrmann et al. (2014) and Manafifard et al. (2016) for evaluating the player detection or tracking results. ISSIA has also been applied for player tracking evaluation (D'Orazio et al., 2009; Martín and Martínez, 2013; Baysal and Duygulu, 2016; Herrmann et al., 2014; Ben Shitrit et al., 2014). Accordingly, integrated trajectories from different cameras of ISSIA were evaluated by D'Orazio et al. (2009) and Martín and Martínez (2013). In Pettersen et al. (2014), a dataset of body-sensor traces and corresponding videos from several soccer games captured in 2013 at the Alfheim Stadium in Norway were presented. Player data (i.e. field position, heading, and speed) were sampled using the highly accurate ZXY sport tracking system.

**Table 5**
A review of occlusion resolution approaches.

| Ref. | Occlusion resolution method | |
|---|---|---|
| | Among teammates | Among competitors etc |
| Iwase and Saito (2002) | Multiple cameras | |
| Iwase and Saito (2003) | Multiple cameras | |
| Choi and Seo (2011) | Multiple cameras | |
| Misu et al. (2009) | Face and back-number recognition | |
| Barceló et al. (2005) | Split-merge in graph | |
| Duh et al. (2013) and Najafzadeh et al. (2015) | Split-merge | |
| Sato and Aggarwal (2005) | Temporal spatio-velocity and computational window deformation | |
| Xu et al. (2004a) | Multiple cameras and prediction using partial observation | |
| Seo et al. (1997) | - | Histogram backprojection |
| Chiang et al. (2009) | Motion prediction | Improved meanshift based on the probability of occluding and occluded player's window |
| Pallavi et al. (2008) | Re-track the mistracked players based on the deviation of the trajectory in graph representation | Correlation of player regions in two consecutive frames using graph representation |
| Khatoonabadi and Rahmati (2009) | Assuming constant speed and direction | Histogram back-projection (occlusions among players and television logos or advertisement boards), split-merge (large occlusion among players), template matching (small occlusion among players) |
| D'Orazio et al. (2009) | - | Searching around predicted position for the best color match |
| Ok et al. (2002) | Reweighting particles in PF using occlusion alarm probability | - |
| Vandenbroucke et al. (1997a) | - | Color classification of the pixels |
| Ohno et al. (2000, 1999) | Positional information | Searching for colors and using color, vertical position and the velocity of players in the case of occlusion among more than two players |
| Intille and Bobick (1995a) | - | Tracking distinguishing intensities |
| Bebie and Bieri (1998) | - | Template matching |
| Figueroa et al. (2006) | Graph: morphological operators, model fitting, split in forward and backward directions, prediction for some short occlusions, trajectories direction and mean velocities of players in particular for occlusion among teammates | |
| Manafifard et al. (2015) | Neighborhood graph: number of parents and children for each node, size of neighboring bounding box, blob dimension and color features | |
| Sullivan and Carlsson (2006) | Graph: constant velocity motion model, fitting the image data and depth ordering | |
| Baysal and Duygulu (2016) | Improved PF | |

Additional player statistics (e.g. total distance covered) were also included. The videos were captured using two stationary camera arrays positioned close to the center of the field.

### 8.2. Playfield detection and player labeling evaluation

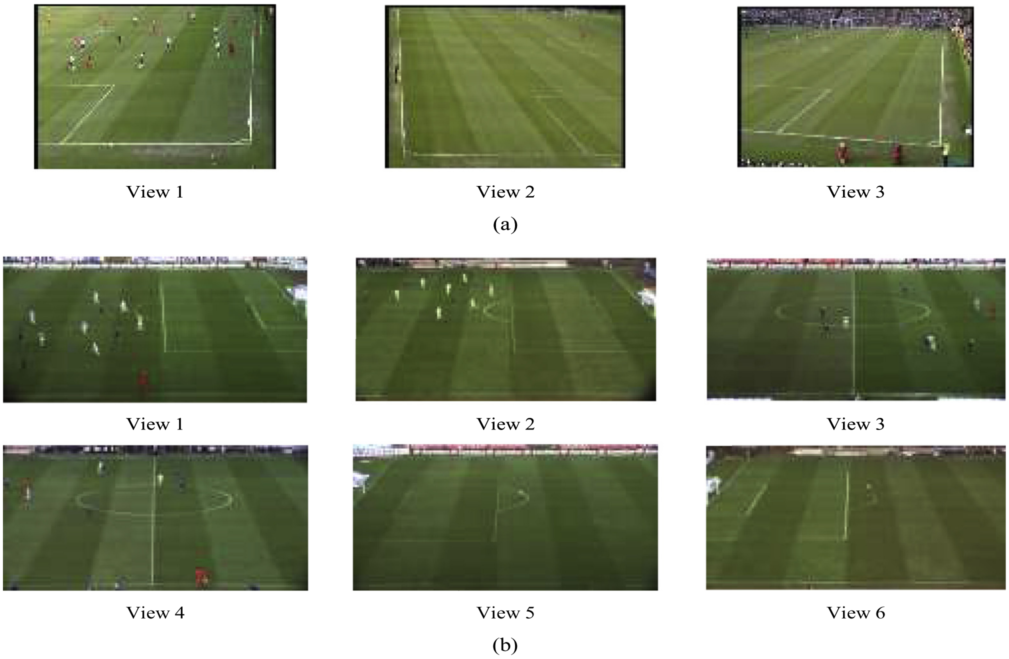
Evaluating the playfield detection was usually ignored except for few works which applied measures including segment precision (true positive /annotated playfield pixels) (Ngo et al., 2010), segment confusion (true negative /annotated non-playfield pixels) (Jiang et al., 2004), the ratio of the number of misclassified pixels versus the total number of pixels (Ekin and Tekalp, 2003) and F-score (Hoernig et al., 2015). In addition to segmentation accuracy, the performance was evaluated in terms of over and under-segmentation by Hung et al. (2011). The performance of the player labeling algorithms was also presented by few works in terms of the percentage of correct classifications (D'Orazio et al., 2009), confusion matrix, detection rate and false alarm rate (Spagnolo et al., 2007). In Baysal and Duygulu (2016), accuracy of labeling was reported with respect to k when k-nearest neighbors algorithm (KNN) leave-one-out cross-validation was applied.

### 8.3. Player detection evaluation

In order to evaluate the player detection step, precision (positive predictivity, specificity) and recall (sensitivity) have been widely applied (Mentzelopoulos et al., 2012; Pallavi et al., 2008; Mackowiak et al., 2010; Utsumi et al., 2002; Khatoonabadi and Rahmati, 2009; Manafifard et al., 2015; Davis, 2008; Tong et al., 2011; Mackowiak and Konieczny, 2012; Manafifard et al., 2016; Sun and Liu, 2009; Tran et al., 2012; Gerke et al., 2013; Ekin and Tekalp, 2003). True positive (TP), true negative (TN), false positive (FP), false negative (FN) (Heydari and Moghadam, 2012; Khatoonabadi and Rahmati, 2009) and the average number of false positives per image (FPPI) (Gerke et al., 2013) have also been presented by some authors. In Mackowiak (2013, 2010), TP, FP and FN were defined by thresholding the degree of overlap between bounding boxes of detected and ground truth regions compared to their union. The receiver operating characteristic (ROC) curves (Xing et al., 2011) and precision-recall curve (Gerke et al., 2013) have also been depicted for evaluation. In Gerke et al. (2013), detection was true positive when it overlapped the ground truth by at least 50%. Relative distance, cover, overlap (Tong et al., 2011) and F-score (Tong et al., 2011; Gerke et al., 2013) have also been used for evaluation. The method by D'Orazio et al. (2009) presented the result of player detection with respect to the ground truth in terms of mean error and variance in pixels and percentage of FP and FN which indicates the amount of players oversegmentation. Missed ratio was also presented as the percentage of the undetected players for a given overlap degree by Mackowiak (2013). Moreover, the quality of the system was measured using a criterion defined as the combination of the precision and missed ratio. The results of the color-based player segmentation were also denoted by the confusion matrix in Vandenbroucke et al. (1998). Moreover, the metric in Renno et al. (2004) computed the accuracy of the segmentation after removing identified shadows by signal to noise ratio.

### 8.4. Player tracking and occlusion handling evaluation

In order to evaluate the player tracking step, multiple object tracking precision (MOTP) and multiple object tracking ac-



View 1

View 2

View 3

(a)

View 1

View 2

View 3

View 4

View 5

View 6

(b)

**Fig. 10.** Sample frames from each view of VS-PETS 2003 and ISSIA dataset. a) VS-PETS 2003 (University of Reading 2016), b) ISSIA (D'Orazio et al., 2009).

curacy (MOTA) (Herrmann et al., 2014), global multiple object tracking accuracy (GMOTA) (Baysal and Duygulu, 2016; Ben Shitrit et al., 2014), global identity mismatch (GMME), FP and FN (Baysal and Duygulu, 2016), track fragmentation and track detection rate (D'Orazio et al., 2009), number of mostly tracked trajectories, partially tracked trajectories, mostly lost trajectories, identity switches and fragmentations of trajectories (Xing et al., 2011), hit and miss measurements (Tong et al., 2011) have been applied. In addition, recall and precision (Martín and Martínez, 2013; Chiang et al., 2009; Manafifard et al., 2015; Davis, 2008; Tong et al., 2011), tracking accuracy (Itoh et al., 2012), reliability of tracking (Li and Flierl, 2012), success and fail rate percentage (Miura and Kubo, 2008; Matsui et al., 1998), detection error defined as sum of FP and FN (Khan and Shah, 2009), number of correctly tracked players (TP) and accuracy (Heydari and Moghadam, 2012), number of tracked frames for indicating persistency of each blob and discarding FP regarding its lifetimes (D'Orazio et al., 2007 ), number of mistrackings and successfully tracked occlusions (Sato and Aggarwal, 2005) have been applied.

A different approach was to evaluate player detection or tracking using position deviation from the ground truth in the image (Xing et al., 2011; Davis, 2008) or model space (Li et al., 2005), which was the most suitable way for evaluation. For this purpose, the error for each track can be calculated as the Euclidean distance from the ground truth (Abbott and Williams, 2007). In Khan and Shah (2009), a tracking accuracy measure was defined as the perpendicular distance between the ground truth line connecting the player head and feet and the least square line through player centroids on all planes (from the player feet to the player head). Then, it was converted to actual world distances. Afterwards, the total average track error was calculated, and it was averaged over the number of players and views. Furthermore, detection error was reported for different number of views with respect to time. Since it was inconvenient to determine the ground truth in multi-camera system with multiple players, the player ground truth was generated only for one view of one particular sequence by D'Orazio et al. (2009). One other approach was to evaluate player tracking from multiple fixed cameras covering the whole pitch by counting the number of tracked players (Xu et al., 2004a). In Li et al. (2005), the spatial accuracy of the tracker was characterized as the proportion of the correctly represented players within Δd meters of the ground truth. The temporal accuracy was also characterized as the proportion of tracks for which its relationship with the ground truth track was maintained. In Needham and Boyle (2001), trajectories within one meter of the ground truth were accepted, since the midpoint of the base of the player bounding box was often considered as the player position. The method by Abbott and Williams (2007) defined the failure as a deviation of at least ten feet from the track's actual location, and then the mean time to failure for individual tracks was computed. Moreover, the root mean square errors of accurately segmented legs (about 0.4 m in average), missed legs under the knees (more than two meters) and only segmented torsos (more than five meters) were presented by Kim et al. (2003), which indicated that the missed feet impose a large amount of deviation in real player position. The main drawback of evaluation on the field model was the dependence of the accuracy evaluation on the accuracy of camera calibration.

In order to evaluate occlusion handling results, accuracy as the ratio of detected occlusions to all occlusions (Pallavi et al., 2008), recall and precision with and without occlusion (Utsumi et al., 2002), percentage of distinguished occluding blobs (D'Orazio et al., 2009 ) and success of tracking over the duration of occlusion events (Choi and Seo, 2011) were applied. To sum up, it is necessary to estimate the performance of each step and then to evaluate the system as a whole, which were neglected by most previous works except for Nunez et al. (2008) in which the player detection was evaluated by the weighted sum of accuracies for its different steps.

## 9. Conclusion and future directions

In the past decades, extensive research has been devoted to soccer player tracking. Generally, precision, robustness, adaptivity, automation and online analysis should be considered to build a robust tracker, which is not an easy task. Since player tracking is often integrated with other preprocessing steps, such as playfield detection, player detection, player labeling and appearance modeling, the tracking results significantly depend on all previous steps. The main goal of this paper was to review different preprocessing steps for soccer player tracking, categorize different tracking frameworks and compare them in terms of the available evaluation criteria. It may help researchers to get familiar with the renowned and state-of-the-art methods in the domain. Moreover, it highlights future research directions to compensate for the weaknesses and low performances of the available algorithms. This paper also gives insight into enhancing existing trackers or proposing new ones, and each reviewed tracker (e.g. PF, graph) can be developed in different applications.

Soccer videos can be categorized into two main categories, namely, videos captured by *i*) stationary and *ii*) moving cameras. Much work has already been done on the latter case including broadcast streams. In broadcast videos, the camera(s) position was often adjusted according to the stadium structure and remained fixed in different matches at the same stadium. Since the broadcast streams were often recorded from broadcast television programs, the camera position was not often reported. Moreover, broadcast videos were affected by different problems, such as small FOV, lack of 3D information, low resolution and complex occlusions. As a solution, multiple stationary cameras were used which required more hardware and computational cost to process large broadcast videos. Moreover, soccer videos were captured by cameras at different frame rates and resolutions. Although high resolution images (e.g. $1920 \times 1080$) were beneficial for tracking and occlusion resolution, medium (e.g. $720 \times 576$) or low (e.g. $352 \times 288$) resolution images were often used; since they required less computational time. A better resolution can also be obtained using multiple cameras with small FOV. Accordingly, soccer videos in multi-view configurations have been captured by a minimum number of cameras to cover the field depending on factors, such as camera resolution, height and FOV. The size of the players and the extent of the pitch visible in the shot were also affected by the camera angle, height and position. However, the camera setup properties in multi-view configurations were rarely reported in the literature. The position of the cameras was also adjusted regarding the layout of the stadium to achieve the best coverage of the field and the best resolution of each area. Moreover, the number of cameras and their FOV, position, height and resolution must be selected with respect to each other. A promising future direction is to investigate the camera setup properties (e.g. camera's height, distance from the field and from other cameras and number of cameras) in order to achieve the best configuration in terms of tracking performance and minimum cost. Moreover, the camera setup properties in multi-view configurations must be reported in more detail by future works.

Challenges for playfield detection are related to the removing various shadows from each light source at the base of a player in some hard sequences and also the players in green uniforms which might be segmented as the green field. Moreover, methods based on training sets and multi-thresholding techniques present limitations due to their sensitivity to parameters adjustment and the change in field color during a game or from one game to another game. Moreover, playfield detection using GMM has outperformed histogram-based methods. Providing more challenging datasets which include shadows seems indispensable for realistic evaluation of playfield detectors.

The quality of player detection is really crucial, since it significantly affects further analysis such as trajectory extraction. Background subtraction in the case of static cameras and classifiers in the case of moving cameras are the two most effective detectors in the literature; however, they relied on binary classification. In order to solve this problem, a blob-guided PSO method was proposed in our previous work. Generally, noisy segmentation of the players and fragmented player regions provide incomplete observations. It is also more difficult to correctly detect the players when their blobs are merged with field marks or advertising billboards. Moreover, player detection may fail in some cases (e.g. fast camera operations) that players appear faded in the scene. The situation is also harder for the distant players who appear blurred and small. Moreover, some player pixels look like green field. On the other hand, most authors have tried to alleviate player detection by first extracting the field area, which can be violated in the case of players in green uniforms and players standing at the border of the field. Therefore, further research should be focused on players at the border of the field to evaluate their detection and tracking performance separately. The main interest feature of a player is also the position of the feet which must be determined with the greatest accuracy for player localization and tracking. However, the player feet may get separated from his body due to the segmentation errors, or the lower leg parts may not be extracted due to factors, such as motion blur, thin legs and small size of distant players. As a result, missed feet may significantly delocalize players on the field model. Missed feet may also cause finding no corresponding blob or wrong corresponding blob during tracking or corresponding players among multiple views. Accordingly, morphological operations can be used for joining the separated feet to the player body, which may cause the neighboring players to merge. Moreover, considering the midpoint of the base of the player bounding box as the feet position is not precise, since the player feet are placed in different distances from the camera. Unfortunately, player detection focusing on accurate segmentation of player feet was ignored in most previous works, which also requires metric player localization assessment. The above issue is fundamental in some applications such as automatic offside detection. Therefore, another promising future direction is player detection and tracking focusing on accurate segmentation of the player feet. Furthermore, background subtraction methods based on motion descriptor are not the best solution for moving cameras. Camera motion compensation also requires more computational cost, and each error involved during the motion compensation directly affects the player detection results. As a solution, uniform color can be used for broadcast sequences due to the discriminative color of the players' uniforms, but detecting distant and blurred players using just color information is an uncertain task. Moreover, the low resolution of pixels and the change in lighting conditions complicate the reliable use of color information. The other challenge is updating the appearance models to avoid the player model deviating from the correct representation. One other limitation of color descriptor is that the color of player uniform must differ from the color of playfield and lines. Moreover, the color model for the player close to the camera and distant from the camera may be different, which need to be investigated. Few papers also focused on combining features for detection improvement. In our opinion, further research must be directed towards developing more promising schemes for combining features to enhance the detection results (e.g. by keeping more feet). Another issue which needs to be investigated is the automatic selection of the appropriate color features without any manual intervention; since it extremely depends on the color of player uniform which changes from one game to another. Furthermore, unsupervised initialization of color models in the first frame of the sequence, in particular in severe occlusion situations, remains an unanswered question. Although few unsupervised player detection methods (e.g. clustering schemes) have been proposed, the robustness, reliability and capability for precise localization have not been guaranteed yet. Therefore, one promising future direction is to improve unsupervised player detection. Generally, most previous player detectors (e.g., detectors based on classifiers and background subtraction) relied on binary classification. Accordingly, they decided if a blob corresponded to the player region or not without locating players in a partially occluding blob. Therefore, the separation of occluding players was mainly sent to the tracking step. However, providing the players' locations in occluding blobs is beneficial for initializing the trackers and also improving player detection from multiple cameras. It is also beneficial for tracking-by-detection algorithms. Although pixel-based classifiers can deal with partial occlusions, the image of a player may get fragmented into multiple regions, and some pixels may be misclassified. It is clear that pixel-based classifiers fail to discriminate lines and players in similar color and also players in same color uniforms. In our opinion, a great deal of work must be directed towards the more promising schemes for detecting partially occluded players and improving pixel-based classifiers. Unsupervised labeling approaches must also be improved to automatically collect the samples from each team uniforms even in the case of occlusions and presence of multiple similar colors in the clothes of competitors. To sum up, improving player detection by focusing on the above challenges is one of the promising future directions.

Player tracking faces difficulty in the cases, such as occluding teammates with abrupt movements, blur, divided player blob into fragments, heavy occlusions, partially visible players at the image borders and newly arrived players in an occluding blob. The player tracking methods were very diverse, and most of them were not totally ideal. Point trackers performed well through partial occlusions, but it might be hard to detect and match interest points for distant or blurred players. The limitation of snakes was their sensitivity to parameters, contour initialization, occlusion or a nonsmooth shape varying process. Although graph provided a beneficial tool for occlusion resolution, the number of look forward or backward frames should be increased in the case of long term occlusions, and it could not be applied for real-time applications. Moreover, nonlinear and unpredictable players' movements might be problematic for KF. As a solution, PF has been applied. The main drawback of PF was its dependence on the number of particles, and the precision of player localization by PF need to be improved. An obvious advantage of meanshift over the standard template matching was avoiding brute force search. However, it required that a portion of the player be inside the initial search window. Since meanshift used a fixed size tracking window, camshift was proposed for size adaptation. Another limitation of some trackers in the literature was ignoring association or assuming the NN strategy as effective; however, ambiguities arose in the case of closely spaced players and high number of false measurements. The graph representation implicitly retained association, but it relied on an ad hoc technique. The main limitation of the original JPDA was its inability to perform track initiation and deletion, and it was appropriate when the number of tracks was known and remained fixed. Moreover, MHT led to large computation and memory resources, which was problematic for real-time applications. The MCMCDA relaxed from one to one correspondence between observations and players. However, detection-based trackers gave poor performance, since detection was not reliable. The main challenge in the literature was to improve different tracking methods and compensate for their weaknesses (e.g. ignoring occlusion, data association and appearance cue) for the specific application of soccer player tracking. The other challenges were to optimize speed, compensate for camera movements in videos captured by moving cameras or using stationary cameras as a solution. Accordingly, information fusion from multiple cameras and different camera configurations were proposed.

Occlusions and nonlinear movements of players and camera were the two main obstacles for soccer player tracking methods. Occlusion resolution by graph and tracking nonlinear movements by PF made them popular in the literature to compensate for weaknesses of soccer player trackers. Accordingly, MCNF and TMCNF, as developed ideas in graph, and Sentioscope, as an advanced version of PF, were the two remarkable player trackers in soccer sequences captured by static cameras. MCNF and TMCNF have been commercialized in different sports (e.g. soccer, volleyball and basketball). They worked on a batch of frames which caused a delay in response depending on the batch size. The longer the batch was, the better response would be resulted. They also outperformed KSP, C-KSP and DP in preserving identities. In a different manner, Sentioscope, as a real-time soccer player tracking software, was conducted on ten full-length soccer matches. It outperformed meanshift, optical flow, color-based PF and color-based mixture PF, KSP, DP and TMCNF. Other tracking methods, such as KF, JPDA, MHT, TSV, contour tracking, MCMCDA, meanshift and camshaft, demonstrated significant weaknesses for soccer player tracking due to specific challenges, such as low resolution image sequences, occlusions and unpredictable relative movements of players and camera. Accordingly, they were occasionally improved in the literature to compensate for their weaknesses. As discussed earlier, PF outperformed KF, and hierarchical PF outperformed meanshift and meanshift-PF. Moreover, improving appearance model in PF based on SMOG outperformed meanshift and condensation based on color histogram. In addition, comparing color-based PF, meanshift and Kalman-meanshift showed that although the latter reduced meanshift iterations, it failed during nonlinear movements. Contrastingly, nonlinear movements were dealt by PF. However, a more precise localization was achieved by the other two trackers. In our opinion, a great deal of work must be directed towards the more promising schemes to compensate for the weaknesses of each tracker. Although these enhancements have been considered in some previous works, all weaknesses were not properly handled, and there are still open research problems. Moreover, tracking methods were combined by few previous works. Accordingly, further research must be directed towards more promising schemes for combining trackers such that the weaknesses of one are compensated by the strengths of the other. For instance, the point tracker can be combined by other trackers regarding its strength for resolving partial occlusions, or the contour tracker can be combined by other trackers regarding its strength for tracking the player contour. The graph representation can also be combined by many trackers due to its strength in resolving occlusions. Generally, computational time decreases by the increase in frame sampling interval at the cost of the decrease in performance. However, the sensitivity of different trackers to the sampling interval differs. For instance, the sensitivity of KF is more than tracking-by-detection methods due to the linear movement assumption. Therefore, one promising future direction is to evaluate the performance of different trackers using different frame sampling intervals.

Occlusion is a primary cause of the tracking failures. Moreover, violations of assumed motion model misled the tracking when occlusions were resolved regarding predictions. Defining a motion model in the case of occluding teammates may also be an uncertain task due to the unpredictable players' movements. Generally, few works have focused on resolving occlusion among more than two players. The performance also decreased when the number of merged players increased or severe occlusion was combined with sudden change in the velocity of players. Moreover, the positioning of the player was quite random in the case of heavily occluded players. Although the need for more cameras is inevitable for high level analysis, advanced image processing methods need to be investigated for analyzing the current broadcast sequences and even improving multi-camera results. Another limitation of existing tracking schemes is ignoring automatic localization of occluding players in the first frame, which needs to be addressed by future works. Moreover, unsupervised initialization of the appearance models, in particular when high proportion of partial occlusions occur or similar colors exist in the competitors' uniforms, is still an unsolved problem. Furthermore, isolating the occluding players in an occluding blob is another crucial factor which influences the player localization precision up to several meters. Assuming the same position for players in an occluding blob or tracking the visible portion of the occluded player drifts the occluding player position. One promising future direction is to focus on the precise localization of the occluding players in an occluding blob in order to reduce the drift in localization.

On the other hand, quantitative evaluation has been ignored by some previous works, or it has been performed via few visual demonstrations or small dataset. It is also difficult to compare methods due to the variability in datasets particularly for broadcast sequences, lack of access to the earlier methods' codes, different evaluation criteria and lack of evaluation process in the literature. Although some datasets and evaluation measures were proposed, the development of a common evaluation measure and dataset, which covers all kinds of scenarios, is still a mandatory task. One promising future direction is to provide soccer broadcast datasets for evaluating methods using common evaluation criteria (e.g. recall, precision, FP, FN, TP, MOTA and GMOTA). Accordingly, MOTA and GMOTA were two effective metrics by considering the switch in players' identities. Another future direction is to propose challenging datasets in order to evaluate occlusion resolution. Moreover, the tracking results for occluding players must be assessed independently (e.g. percentage of resolved occlusions between two or more), since a large number of non-occluding players significantly improve the overall result. Different steps of each paper can also influence the next steps, and the variability in applied steps complicates evaluation of the final step (e.g. tracking). Therefore, it is necessary to estimate the performance of each step and then evaluate the system as a whole, which was neglected by most previous works. Moreover, some datasets must be proposed for playfield detection. In addition, the evaluations of the playfield detection and player labeling were insufficient or neglected by most authors. More precise criteria, such as metric evaluation of the player localization, are also needed to be proposed. However, several factors, such as the accuracy of the calibration data in the case of metric evaluation and the accuracy of the ground truth (errors in marking the ground truth in particular for heavily occluded players are inevitable), still affect the accuracy of the evaluation. In conclusion, improvements in each prior step reviewed earlier will certainly lead to more precise player tracking and high level soccer video analysis. We hope that this article gives a valuable insight into this research topic and leads to new solutions to the challenges it is facing.

## References

Abbott, R.G., Williams, L.R., 2007. Multiple target tracking with lazy background subtraction and connected components analysis. Mach. Vis. Appl. 20, 93–101.

Arulampalam, M.S., Maskell, S., Gordon, N., Clapp, T., 2002. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Trans. Signal Process. 50, 174–188.

Assfalg, J., Bertini, M., Colombo, C., Del Bimbo, A., Nunziati, W., 2003. Semantic annotation of soccer videos: automatic highlights identification. Comput. Vis. Image Understand. 92, 285–305.

Avidan, S., 2007. Ensemble tracking. IEEE Trans. Pattern Anal. Mach. Intell. 29, 261–271.

Bai, X., Zhang, T., Song, X., Niu, X., 2011. Playfield detection using color ratio and local entropy. In: Proc. 7th Int. Conf. on Intelligent Information Hiding and Multimedia Signal Processing, pp. 356–359.

Barceló, L., Binefa, X., Kender, J.R., 2005. Robust methods and representations for soccer player tracking and collision resolution. In: Proc. 4th Int. Conf. on Image and Video Retrieval, Singapore, July 20-22, pp. 237–246.

Barnard, M., Odobez, J., 2004. Robust playfield segmentation using MAP adaptation. In: Proc. 17th Int. Conf. on Pattern Recognition, pp. 610–613.

Barros, R.M.L., Misuta, M.S., Menezes, R.P., Figueroa, P.J., Moura, F.A., Cunha, S.A., Anido, R., Leite, N.J., 2007. Analysis of the distances covered by first division brazilian soccer players obtained with an automatic tracking method. J. Sports Sci. Med. 6, 233–242.

Baysal, S., Duygulu, P., 2016. Sentioscope: a soccer player tracking system using model field particles. IEEE Trans. Circuits Syst. Video Technol. 26, 1350–1362.

Bebie, T., Bieri, H., 1998. SoccerMan-reconstructing soccer games from video sequences. In: Proc. Int. Conf. on Image Processing, pp. 898–902.

Beetz, M., Gedikli, S., Bandouch, J., Kirchlechner, B., Hoyningen-Huene, N.V., Perzylo, A., 2007. Visually tracking football games based on TV broadcasts. In: Proc. 20th Int. Joint Conf. on Artifical intelligence. Hyderabad, India, pp. 2066–2071.

Beetz, M., Hoyningen-Huene, N.V., Bandouch, J., Kirchlechner, B., Gedikli, S., Maldonado, A., 2006. Camera-based observation of football games for analyzing multi--agent activities. In: Proc. 5th Int. Joint Conf. on Autonomous Agents and Multiagent Systems Hakodate. Japan, pp. 42–49.

Ben Shitrit, H., Berclaz, J., Fleuret, F., Fua, P., 2014. Multi-commodity network flow for tracking multiple people. IEEE Trans. Pattern Anal. Mach. Intell. 36, 1614–1627.

Blackman, S.S., 2004. Multiple hypothesis tracking for multiple target tracking. IEEE Aerosp. Electron. Syst. Mag. 19, 5–18.

Bradski, G.R., 1998. Computer vision face tracking for use in a perceptual user interface. Intel Technol. J. 1–15.

Bu, J., Lao, S., Bai, L., 2011. Automatic line mark recognition and its application in camera calibration in soccer video. In: Proc. IEEE Int. Conf. on Multimedia and Expo, pp. 1–6.

Capturing and visualizing large scale human action (accessed 16.09 (accessed 16.09.

Chai, Y., Park, J., Yoon, K., Kim, T., 2011. Multi target tracking using multiple independent particle filters for video surveillance. In: Proc. IEEE Int. Conf. on Consumer Electronics, pp. 735–736.

Chiang, T.-K., Leou, J.-J., Lin, C.-S., 2009. An improved mean shift algorithm based tracking system for soccer game analysis. In: Proc. Asia-Pacific Signal and Information Processing Association. Sapporo, Japan, pp. 380–385.

Choi, K., Seo, Y., 2011. Automatic initialization for 3D soccer player tracking. Pattern Recognit. Lett. 32, 1274–1282.

Choi, K., Seo, Y., Lee, S.W., 2004. Probabilistic tracking of soccer players and ball. In: Proc. Asian Conf. on Computer Vision.

Cox, I.J., 1993. A review of statistical data association techniques for motion correspondence. Int. J. Comput. Vis. 10, 53–66.

Cox, I.J., Hingorani, S.L., 1996. An efficient implementation of Reid's multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking. IEEE Trans. Pattern Anal. Mach. Intell. 18, 138–150.

Cox, I.J., Leonard, J.J., 1991. Probabilistic data association for dynamic world modeling: a multiple hypothesis approach. In: Proc. 5th Int. Conf. on Advanced Robotics. Robots in Unstructured Environments, pp. 1287–1294.

D'Orazio, T., Leo, M., 2010. A review of vision-based systems for soccer video analysis. Pattern Recognit. 43, 2911–2926.

D'Orazio, T., Leo, M., Mosca, N., Spagnolo, P., Mazzeo, P.L., 2009. A semi-automatic system for ground truth generation of soccer video sequences. 6th IEEE Int. Conf. on Advanced Video and Signal Surveillance http://www.ino.it/home/spagnolo/Dataset.html. (accessed: 17.06.2014).

Davis, M., 2008. Investigation into tracking football players from single viewpoint video sequences. Bachelor of Science in Computer Science With Honours. The University of Bath.

Dearden, A., Demiris, Y., Grau, O., 2006. Tracking football player movement from a single moving camera. In: Proc. 3th European Conf. on Visual Media Production, pp. 29–37.

D'Orazio, T., Leo, M., Spagnolo, P., Mazzeo, P.L., Mosca, N., Nitti, M., 2007. A visual tracking algorithm for real time people detection. 8th Int. Workshop on Image Analysis for Multimedia Interactive Services 34-34.

D'Orazio, T., Leo, M., Spagnolo, P., Mazzeo, P.L., Mosca, N., Nitti, M., Distante, A., 2009. An investigation into the feasibility of real-time soccer offside detection from a multiple camera system. IEEE Trans. Circuits Syst. Video Technol. 19, 1804–1818.

Du, W., Hayet, J.-b., Piater, J., Verly, J., 2006. Collaborative multi-camera tracking of athletes in team sport. In: Workshop on Computer Vision Based Analysis in Sport Environments (CVBASE), pp. 2–13.

Du, W., Piater, J., 2007. Multi-camera people tracking by collaborative particle filters and principal axis-based integration. In: Proc. 8th Asian Conf. on Computer Vision. Tokyo, Japan, pp. 365–374. November 18-22.

Duh, D.-J., Chang, S.-Y., Chen, S.-Y., Kan, C.-C., 2013. Automatic broadcast soccer video analysis, player detection, and tracking based on color histogram. Intell. Technol. Eng. Syst. 123–130.

Ekin, A., Tekalp, A.M., 2003. Robust dominant color region detection and color-based applications for sports video. In: Proc. Int. Conf. on Image Processing, pp. 21–24.

Ekin, A., Tekalp, A.M., Mehrotra, R., 2003. Automatic soccer video analysis and summarization. IEEE Trans. Image Process. 12, 796–807.

Enomoto, A., Saito, H., 2009. AR display for observing sports events based on camera tracking using pattern of ground. In: Proc. 3th Int. Conf. on Virtual and Mixed Reality, San Diego, CA, USA, July 19-24, pp. 421–430.

Figueroa, P., Leite, N., Barros, R.M.L., Cohen, I., Medioni, G., 2004. Tracking soccer players using the graph representation. In: Proc. 17th Int. Conf. on Pattern Recognition, pp. 787–790.

Figueroa, P.J., Leite, N.J., Barros, R.M.L., 2006. Tracking soccer players aiming their kinematical motion analysis. Comput. Vis. Image Understand. 101, 122–135.

Fortmann, T.E., Bar-Shalom, Y., Scheffe, M., 1983. Sonar tracking of multiple targets using joint probabilistic data association. IEEE J. Oceanic Eng. 8, 173–184.

Fukunaga, K., Hostetler, L., 1975. The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Trans. Inf. Theory 21, 32–40.

Gabriel, P., Hayet, J.B., Piater, J., Verly, J., 2005. Object tracking using color interest points. In: Proc. IEEE Conf. on Advanced Video and Signal Based Surveillance, pp. 159–164.

Gedikli, S., Bandouch, J., von Hoyningen-Huene, N., Kirchlechner, B., Beetz, M., 2007. An adaptive vision system for tracking soccer players from variable camera settings. In: Proc. 5th Int. Conf. on Computer Vision Systems.

Gerke, S., Singh, S., Linnemann, A., Ndjiki-Nya, P., 2013. Unsupervised color classifier training for soccer player detection. In: Proc. Int. Conf. on Visual Communications and Image Processing, pp. 1–5.

Gudmundsson, J., Horton, M., 2016. Spatio-Temporal Analysis of Team Sports-A Survey, pp. 1–42 http://arxiv.org/abs/1602.06994, (accessed 16.09.

Hamid, R., Kumar, R.K., Grundmann, M., Kihwan, K., Essa, I., Hodgins, J., 2010. Player localization using multiple static cameras for sports visualization. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 731–738.

Hashimoto, S., Ozawa, S., 2006. A system for automatic judgment of offsides in soccer games. In: Proc. IEEE Int. Conf. on Multimedia and Expo, pp. 1889–1892.

Hayet, J.B., Mathes, T., Czyz, J., Piater, J., Verly, J., Macq, B., 2005. A modular multi--camera framework for team sports tracking. In: Proc. IEEE Conf. on Advanced Video and Signal Based Surveillance, pp. 493–498.

Herrmann, M., Hoernig, M., Radig, B., 2014. Online multi-player tracking in monocular soccer videos. AASRI Procedia 8, 30–37.

Heydari, M., Moghadam, A.M.E., 2012. An MLP-based player detection and tracking in broadcast soccer video. In: Proc. Int. Conf. on Robotics and Artificial Intelligence, pp. 195–199.

Hoernig, M., Herrmann, M., Radig, B., 2015. Real-time segmentation methods for monocular soccer videos. Pattern Recognit Image Anal. 25, 327–337.

Hoyningen-Huene, N.V., Beetz, M., 2009. Rao-Blackwellized resampling particle filter for real-time player tracking in sports. In: Proc. 4th Int. Conf. on Computer Vision Theory and Applications, Lisboa, Portugal, 1, pp. 464–471.

Huang, Y., Llach, J., Bhagavathy, S., 2007. Players and ball detection in soccer videos based on color segmentation and shape analysis. In: Int. Workshop on Multimedia Content Analysis and Mining, Weihai, China, June 30-July 1, pp. 416–425.

Hung, M.-H., Hsieh, C.-H., Kuo, C.-M., Pan, J.-S., 2011. Generalized playfield segmentation of sport videos using color features. Pattern Recognit. Lett. 32, 987–1000.

Inamoto, N., Saito, H., 2007. Virtual viewpoint replay for a soccer match by view interpolation from multiple cameras. IEEE Trans. Multimedia 9, 1155–1166.

Intille, S.S., Bobick, A.F., 1995b. Closed-world tracking. In: Proc. 5th Int. Conf. on Computer Vision, pp. 672–678.

Intille, S.S., Bobick, A.F., 1995a. Visual tracking using closed-worlds. In: Proc. Int. Conf. on Computer Vision.

Itoh, H., Takiguchi, T., Ariki, Y., 2012. 3D tracking of soccer players using time-situation graph in monocular image sequence. In: Proc. 21th Int. Conf. on Pattern Recognition, pp. 2532–2536.

Iwase, S., Saito, H., 2002. Tracking soccer player using multiple views. In: Int. Association for Pattern Recognition (IAPR) Workshop on Machine Vision Applications, pp. 102–105.

Iwase, S., Saito, H., 2003. Tracking soccer players based on homography among multiple views. Visual Commun. Image Process. 5150, 283–292.

Iwase, S., Saito, H., 2004. Parallel tracking of all soccer players by integrating detected positions in multiple view images. In: Proc. 17th Int. Conf. on Pattern Recognition, pp. 751–754.

Jahandide, H., Mohamedpour, K., Abrishami Moghaddam, H., 2012. A hybrid motion and appearance prediction model for robust visual object tracking. Pattern Recognit. Lett. 33, 2192–2197.

Jiang, S., Ye, Q., Gao, W., Huang, T., 2004. A new method to segment playfield and its applications in match analysis in sports video. In: Proc. 12th Annual Association for Computing Machinery (ACM) Int. Conf. on Multimedia, New York, NY, USA, pp. 292–295.

Joo, S.-W., Chellappa, R., 2007. A multiple-hypothesis approach for multiobject visual tracking. IEEE Trans. Image Process. 16, 2849–2854.

Junior, B.M., Anido, R.d.O., 2004. Distributed real-time soccer tracking, Association for Computing Machinery (ACM) 2nd Int. Workshop on Video Surveillance & Amp; Sensor Networks. New York, NY, USA, pp. 97–103.

Kang, J., Cohen, I., Medioni, G., 2004. Tracking people in crowded scenes across multiple cameras. In: Proc. Asian Conf. on Computer Vision.

Kangarloo, K., Kabir, E., 2004. Grass field segmentation, the first step toward player tacking, deep compression, and content based football image retrieval. In: Proc. Int. Conf. on Image Analysis and Recognition, Porto. Portugal, pp. 818–824.

Kangarloo, K., Kabir, E., 2005. Sequential probabilistic grass field segmentation of soccer video images. In: 10th Int. Workshop on Combinatorial Image Analysis, Auckland, New Zealand, December 1-3, pp. 639–645.

Kasuya, N., Kitahara, I., Kameda, Y., Ohta, Y., 2008. Robust trajectory estimation of soccer players by using two cameras. In: Proc. 19th Int. Conf. on Pattern Recognition, pp. 1–4.

Kawashima, T., Yoshino, L., Aoki, Y., 1994. Qualitative image analysis of group behaviour. In: Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, pp. 690–693.

Kayumbi, G., Anjum, N., Cavallaro, A., 2008. Global trajectory reconstruction from distributed visual sensors. In: Proc. 2th ACM/IEEE Int. Conf. on Distributed Smart Cameras, pp. 1–8.

Khan, S.M., Shah, M., 2009. Tracking multiple occluding people by localizing on multiple scene planes. IEEE Trans. Pattern Anal. Mach. Intell. 31, 505–519.

Khatoonabadi, S.H., Rahmati, M., 2009. Automatic soccer players tracking in goal scenes by camera motion elimination. Image Vis. Comput. 27, 469–479.

Kim, H., Nam, S., Kim, J., 2003. Player segmentation evaluation for trajectory estimation in soccer games. In: Proc. Image and Vision Computing, Palmerston North. New Zealand, pp. 159–162.

Le Troter, A., Mavromatis, S., Sequeira, J., 2004. Soccer field detection in video images using color and spatial coherence. In: Proc. Int. Conf. on Image Analysis and Recognition, Porto, Portugal, September 29 - October 1, pp. 265–272.

Lefèvre, S., Fluck, C., Maillard, B., Vincent, N., 2000. A fast snake-based method to track football players. In: Int. Association for Pattern Recognition (IAPR) Workshop on Machine Vision Applications, pp. 501–504.

Lefèvre, S., Vincent, N., 2004. Real time multiple object tracking based on active contours. In: Proc. Int. Conf. on Image Analysis and Recognition, Porto, Portugal, September 29 - October 1, pp. 606–613.

Li, H., Flierl, M., 2012. Sift-based multi-view cooperative tracking for soccer video. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 1001–1004.

Li, S., Lu, B., 2007. Automatic camera calibration technique and its application in virtual advertisement insertion system. In: Proc. 2th IEEE Int. Conf. on Industrial Electronics and Applications, pp. 288–292.

Li, Y., Dore, A., Orwell, J., 2005. Evaluating the performance of systems for tracking football players and ball. In: Proc. IEEE Conf. on Advanced Video and Signal Based Surveillance, pp. 632–637.

Li, Y., Liu, G., Qian, X., 2009. Ball and field line detection for placed kick refinement. In: World Resources Institute (WRI) Global Congress on Intelligent Systems, pp. 404–407.

Liu, J., Tong, X., Li, W., Wang, T., Zhang, Y., Wang, H., 2009. Automatic player detection, labeling and tracking in broadcast soccer video. Pattern Recognit. Lett. 30, 103–113.

Liu, Y., Guo, M., Liu, W., 2011. Detection of playfield with shadow and its application to player tracking. In: IEEE Int. Workshop on Machine Learning for Signal Processing, pp. 1–5.

Mackowiak, S., 2013. Segmentation of football video broadcast. Int. J. Electron. Telecommun. 59, 75–84.

Mackowiak, S., Konieczny, J., 2012. Player extraction in sports video sequences. In: Proc. 19th Int. Conf. on Systems, Signals and Image Processing, pp. 409–412.

Maćkowiak, S., Konieczny, J., Kurc, M., Maćkowiak, P., 2010. A complex system for football player detection in broadcasted video. In: Proc. Int. Conf. on Signals and Electronic Systems, pp. 119–122.

Manafifard, M., Ebadi, H., Abrishami-Moghaddam, H., 2015. Discrete particle swarm optimization for player trajectory extraction in soccer broadcast videos. Scientia Iranica 22, 1031–1044.

Manafifard, M., Ebadi, H., Moghaddam, H.A., 2016. Multi-player detection in soccer broadcast videos using a blob-guided particle swarm optimization method. Multimedia Tools Appl. 1–30.

Marchesotti, L., Vernazza, G., Regazzoni, C., 2004. A multicamera fusion framework for multiple occluding objects tracking in intelligent monitoring and sport viewing applications. In: Proc. Int. Conf. on Image Processing, pp. 1033–1036.

Martín, R., Martínez, J.M., 2013. A semi-supervised system for players detection and tracking in multi-camera soccer videos. Multimedia Tools Appl. 73, 1617–1642.

Mathes, T., Piater, J.H., 2006. Robust non-rigid object tracking using point distribution manifolds. In: 28th DAGM Symp. on Pattern Recognition, Berlin, Germany, September 12-14, pp. 515–524.

Matsui, K., Iwase, M., Agata, M., Tanaka, T., Ohnishi, N., 1998. Soccer image sequence computed by a virtual camera. In: Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, pp. 860–865.

Mazzeo, P.L., Spagnolo, P., Leo, M., D'Orazio, T., 2008. Visual players detection and tracking in soccer matches. In: Proc. 5th IEEE Int. Conf. on Advanced Video and Signal Based Surveillance, pp. 326–333.

Mentzelopoulos, M., Psarrou, A., Angelopoulou, A., García-Rodríguez, J., 2012. Active foreground region extraction and tracking for sports video annotation. Neural Process. Lett. 37, 33–46.

Misu, T., Gohshi, S., Izumi, Y., Fujita, Y., Naemura, M., 2004. Robust tracking of athletes using multiple features of multiple views. In: Proc. Int. Conf. in Central Europe on Computer Graphics, Visualization and Computer Vision, Plzen-Bory. Czech Republic, pp. 285–292.

Misu, T., Matsui, A., Clippingdale, S., Fujii, M., Yagi, N., 2009. Probabilistic integration of tracking and recognition of soccer players. In: Proc. 15th Int. Conf. on Multimedia Modeling, Sophia-Antipolis, France, January 7-9, pp. 39–50.

Misu, T., Naemura, M., Wentao, Z., Izumi, Y., Fukui, K., 2002. Robust tracking of soccer players based on data fusion. In: Proc. 16th Int. Conf. on Pattern Recognition, pp. 556–561.

Miura, J., Kubo, H., 2008. Tracking players in highly complex scenes in broadcast soccer video using a constraint satisfaction approach. In: Proc. Int. Conf. on Content-based Image and Video Retrieval. Niagara Falls, Canada, pp. 505–514.

Mochizuki, T., Fujii, M., Shibata, M., Sakai, Y., 2009. Fast identification of player position in soccer broadcast video by block-based camera view angle search. In: 6th Int. Symp. on Image and Signal Processing and Analysis, pp. 408–413.

Montañés Laborda, M.A., Torres Moreno, E.F., Martínez del Rincón, J., Herrero Jaraba, J.E., 2011. Real-time GPU color-based segmentation of football players. J. Real-Time Image Process. 7, 267–279.

Naemura, M., Fukuda, A., Mizutani, Y., Izumi, Y., Tanaka, Y., Enami, K., 2000. Morphological segmentation of sport scenes using color information. IEEE Trans. Broadcast. 46, 181–188.

Naidoo, W.C., Tapamo, J.R., 2006. Soccer video analysis by ball, player and referee tracking. In: Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on IT Research in Developing Countries. Somerset West, South Africa, pp. 51–60.

Najafzadeh, N., Fotouhi, M., Kasaei, S., 2015. Multiple soccer players tracking. In: Int. Symp. on Artificial Intelligence and Signal Processing, pp. 310–315.

Needham, C.J., Boyle, R.D., 2001. Tracking multiple sports players through occlusion, congestion and scale. In: Proc. British Machine Vision Conference, pp. 93–102.

Ngo, V.A., Yang, W., Cai, J., 2010. Accurate playfield detection using area-of-coverage. In: IEEE Int. Symp. on Circuits and Systems, pp. 3441–3444.

Nummiaro, K., Koller-Meier, E., Gool, L.V., 2003. An adaptive color-based particle filter. Image Vis. Comput. 21, 99–110.

Nunez, J.R., Facon, J., De Souza Brito, A., 2008. Soccer video segmentation: referee and player detection. In: Proc. 15th Int. Conf. on Systems, Signals and Image Processing, pp. 279–282.

Oh, S., Russell, S., Sastry, S., 2004. Markov chain Monte Carlo data association for general multiple-target tracking problems. In: Proc. 43th IEEE Conf. on Decision and Control, pp. 735–742.

Ohno, Y., Miura, J., Shirai, Y., 2000. Tracking players and estimation of the 3D position of a ball in soccer games. In: Proc. 15th Int. Conf. on Pattern Recognition, pp. 145–148.

Ohno, Y., Miurs, J., Shirai, Y., 1999. Tracking players and a ball in soccer games. In: Proc. IEEE/SICE/RSJ Int. Conf. on Multisensor Fusion and Integration for Intelligent Systems, pp. 147–152.

Ok, H.W., Seo, Y., Hong, K.S., 2002. Multiple soccer players tracking by condensation with occlusion alarm probability. Int. Workshop on Statistical Methods for Vision Processing.

Oskouie, P., Alipour, S., Eftekhari-Moghadam, A.-M., 2014. Multimodal feature extraction and fusion for semantic mining of soccer video: a survey. Artif. Intell. Rev. 42, 173–210.

Pallavi, V., Mukherjee, J., Majumdar, A.K., Sural, S., 2008. Graph-based multiplayer detection and tracking in broadcast soccer videos. IEEE Trans. Multimedia 10, 794–805.

Pettersen, S.A., Johansen, D., Johansen, H., Berg-Johansen, V., Gaddam, V.R., Mortensen, A., Langseth, R., Griwodz, C., Stensland, H.K., Halvorsen, P., 2014. Soccer video and player position dataset. In: Proc. Int. Conf. on Multimedia Systems. Singapore.

Poppe, C., Bruyne, S.D., Verstockt, S., de Walle, R.V., 2010. Multi-camera analysis of soccer sequences. In: Proc. 7th IEEE Int. Conf. on Advanced Video and Signal Based Surveillance, pp. 26–31.

Qian, H., Mao, Y., Geng, J., Wang, Z., 2007. Object tracking with self-updating tracking window. In: Pacific Asia Workshop on Intelligence and Security Informatics. Chengdu, China, pp. 82–93. April 11-12.

Rangsee, P., Suebsombat, P., Boonyanant, P., 2013. Simplified low-cost GPS-based tracking system for soccer practice. In: Proc.13th Int. Symp. on Communications and Information Technologies, pp. 724–728.

Reid, D.B., 1979. An algorithm for tracking multiple targets. IEEE Trans. Autom. Control 24, 843–854.

Ren, J., Xu, M., Orwell, J., Jones, G.A., 2009. Multi-camera video surveillance for real–time analysis and reconstruction of soccer games. Mach. Vis. Appl. 21, 855–863.

Renno, J.R.R., Orwell, J., Thirde, D.J., Jones, G.A., 2004. Shadow classification and evaluation for soccer player detection. In: Proc. British Machine Vision Conference, pp. 839–848.

Sato, K., Aggarwal, J.K., 2004. Temporal spatio-velocity transform and its application to tracking and interaction. Comput. Vis. Image Understand. 96, 100–128.

Sato, K., Aggarwal, J.K., 2005. Tracking soccer players using broadcast TV images. In: Proc. IEEE Conf. on Advanced Video and Signal Based Surveillance, pp. 546–551.

Schlipsing, M., Salmen, J., Tschentscher, M., Igel, C., 2014. Adaptive pattern recognition in real-time video-based soccer analysis. J. Real-Time Image Process. 1–17.

Seo, Y., Choi, S., Kim, H., Hong, K.-S., 1997. Where are the ball and players? Soccer game analysis with color-based tracking and image mosaick. In: Proc. 9th Int. Conf. on Image Analysis and Processing, Florence, Italy, September 17–19, pp. 196–203.

Spagnolo, P., Mosca, N., Nitti, M., Distante, A., 2007. An unsupervised approach for segmentation and clustering of soccer players. In: Proc. Int. Conf. on Machine Vision and Image Processing, pp. 133–142.

Sullivan, J., Carlsson, S., 2006. Tracking and labelling of interacting multiple targets. In: Proc. 9th European Conf. on Computer Vision, Graz, Austria, May 7-13, pp. 619–632.

Sullivan, J., Nillius, P., Carlsson, S., 2009. Multi-target tracking on a large scale: experiences from football player tracking. In: Proc. IEEE Int. Conf. on Robotics and Automation (ICRA) Workshop on People Detection and Tracking. Kobe, Japan.

Sun, L., Liu, G., 2009. Field lines and players detection and recognition in soccer video. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 1237–1240.

Svensson, D., 2010. Target Tracking in Complex Scenarios Thesis For The Degree Of Doctor Of Philosophy. Chalmers University of Technology, Department of Signals and Systems, Goteborg, Sweden.

Tabii, Y., Thami, R.O.H., 2009. A framework for soccer video processing and analysis based on enhanced algorithm for dominant color extraction. Int. J. Image Process..

Taki, T., Hasegawa, J., Fukumura, T., 1996. Development of motion analysis system for quantitative evaluation of teamwork in soccer games. In: Proc. Int. Conf. on Image Processing, pp. 815–818.

Tong, X., Liu, J., Wang, T., Zhang, Y., 2011. Automatic player labeling, tracking and field registration and trajectory mapping in broadcast soccer video. Assoc. Comput. Mach.y (ACM) Trans. Intell. Syst. Technol. 2, 1–32.

Tran, Q., Tran, A., Dinh, T.B., Duong, D., 2012. Long-view player detection framework algorithm in broadcast soccer videos. In: Proc. 7th Int. Conf. on Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence. Zhengzhou, China, pp. 557–564. August 11-14.

University of Reading, 2003. VS-PETS football dataset. The First Joint IEEE Int. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, Nice, October http://www.cvg.reading.ac.uk/slides/pets.html, (accessed 17.06.2016).

Utsumi, O., Miura, K., Ide, I., Sakai, S., Tanaka, H., 2002. An object detection method for describing soccer games from video. In: Proc. IEEE Int. Conf. on Multimedia and Expo, pp. 45–48.

Vandenbroucke, N., Macaire, L., Postaire, J.G., 1997. Soccer Player Recognition By Pixel Classification in a Hybrid Color Space, pp. 23–33.

Vandenbroucke, N., Macaire, L., Postaire, J.G., 1998. Color pixels classification in an hybrid color space. In: Proc. Int. Conf. on Image Processing, pp. 176–180.

Vandenbroucke, N., Macaire, L., Postaire, J.-G., 2003. Color image segmentation by pixel classification in an adapted hybrid color space. Application to soccer image analysis, Computer Vision and Image Understanding 90, 190–216.

Vandenbroucke, N., Macaire, L., Vieren, C., Postaire, J.G., 1997. Contribution of a color classification to soccer players tracking with snakes. In: Proc. IEEE Int. Conf. on Systems, Man, and Cybernetics, Computational Cybernetics and Simulation, pp. 3660–3665.

Vermaak, J., Doucet, A., Perez, P., 2003. Maintaining multimodality through mixture tracking. In: Proc. 9th IEEE Int. Conf. on Computer Vision, pp. 1110–1116.

Vermaak, J., Maskell, S., Briers, M., 2005. A unifying framework for multi-target tracking and existence. In: Proc. 8th Int. Conf. on Information Fusion.

Viola, P., Jones, M.J., 2004. Robust real-time face detection. Int. J. Comput. Vis. 57, 137–154.

Vos, R., Brink, W., 2009. Combining Motion Detection And Hierarchical Particle Filter Tracking in a Multi-Player Sports Environment www.prasa.org/proceedings/2009/prasa09-12.pdf.

Wang, H., Suter, D., Schindler, K., 2006. Effective appearance model and similarity measure for particle filtering and visual tracking. In: Proc. 9th European Conf. on Computer Vision, Graz, Austria, May 7-13, pp. 606–618.

Wang, L., Zeng, B., Lin, S., Xu, G., Shum, H.-Y., 2004. Automatic extraction of semantic colors in sports video. In: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, pp. 617–620.

Wang, S.-T., Leou, J.-J., Lin, C.-S., 2008. A new hierarchical particle filter based tracking system for soccer game analysis. In: Proc. 9th Pacific Rim Conf. on Multimedia, Tainan, Taiwan, December 9-13, pp. 358–367.

Watanabe, T., Haseyama, M., Kitajima, H., 2004. A soccer field tracking method with wire frame model from TV images. In: Proc. Int. Conf. on Image Processing, pp. 1633–1636.

Wu, Y., Tong, X., Zhang, Y., Lu, H., 2008. Boosted interactively distributed particle filter for automatic multi-object tracking. In: Proc. 15th IEEE Int. Conf. on Image Processing, pp. 1844–1847.

Xing, J., Ai, H., Liu, L., Lao, S., 2011. Multiple player tracking in sports video: a dual–mode two-way bayesian inference approach with progressive observation modeling. IEEE Trans. Image Process. 20, 1652–1667.

Xu, M., Orwell, J., Jones, G., 2004. Tracking football players with multiple cameras. In: Proc. Int. Conf. on Image Processing, pp. 2909–2912.

Xu, M., Orwell, J., Lowey, L., Thirde, D., 2004. Architecture and algorithms for tracking football players with multiple cameras. In: IEE Proc. Vision, Image and Signal Processing, pp. 51–55.

Xu, P., Xie, L., Chang, S.-F., Divakaran, A., Vetro, A., Sun, H., 2001. Algorithms and system for segmentation and structure analysis in soccer video. In: Proc. IEEE Int. Conf. on Multimedia and Expo, pp. 721–724.

Yang, C., Duraiswami, R., Davis, L., 2005. Fast multiple object tracking via a hierarchical particle filter. In: Proc. 10th IEEE Int. Conf. on Computer Vision, pp. 212–219.

Yang, H., Shao, L., Zheng, F., Wang, L., Song, Z., 2011. Recent advances and trends in visual tracking: a review. Neurocomputing 74, 3823–3831.

Yao, A., Uebersax, D., Gall, J., Gool, L., 2010. Tracking people in broadcast sports, 32th DAGM Symp, Darmstadt, Germany, September 22-24, pp. 151–161.

Yilmaz, A., Javed, O., Shah, M., 2006. Object tracking: a survey. Assoc. Comput. Mach. (ACM) J. Comput. Surv. 38, 1–44.

Yoon, H.-S., Bae, Y.-l.J., Yang, Y.-k., 2002. A soccer image sequence mosaicking and analysis method using line and advertisement board detection. Electron. Telecommun. Res. Inst. (ETRI) J. 24, 443–454.

Yu, J., Tang, Y., Wang, Z., Shi, L., 2007. Playfield and ball detection in soccer video. In: 3th Int. Symp. on Advances in Visual Computing, Lake Tahoe, NV, USA, November 26-28, pp. 387–396.

Zhang, Y., Lu, H., Xu, C., 2008. Collaborate ball and player trajectory extraction in broadcast soccer video. In: Proc. 19th Int. Conf. on Pattern Recognition, pp. 1–4.

Zheng, N., Xue, J., 2009. Multi-target tracking in video – Part II, Statistical Learning and Pattern Analysis for Image and Video Processing. London, pp. 319–341.

Zhong, X., Zheng, N., Xue, J., 2006. Pseudo measurement based multiple model approach for robust player tracking. In: Proc. 7th Asian Conf. on Computer Vision, Hyderabad, India, January 13-16, pp. 781–790.

Zhu, G., Huang, Q., Xu, C., Rui, Y., Jiang, S., Gao, W., Yao, H., 2007. Trajectory based event tactics analysis in broadcast sports video. In: Proc. 15th Association for Computing Machinery (ACM) Int. Conf. on MultimediaAugsburg. Germany, pp. 58–67.

Zhu, G., Liang, D., Liu, Y., Huang, Q., Gao, W., 2005. Improving particle filter with support vector regression for efficient visual tracking. In: Proc. IEEE Int. Conf. on Image Processing, pp. 422–425.

Zhu, G., Xu, C., Huang, Q., Gao, W., 2006. Automatic multi-player detection and tracking in broadcast sports video using support vector machine and particle filter. In: Proc. IEEE Int. Conf. on Multimedia and Expo, pp. 1629–1632.

Zhu, G., Xu, C., Huang, Q., Rui, Y., Jiang, S., Gao, W., Yao, H., 2009. Event tactic analysis based on broadcast sports video. IEEE Trans. Multimedia 11, 49–67.