A Holistic Approach for Role Inference and Action Anticipation in Human Teams

JUNYI DONG, QINGZE HUO, and SILVIA FERRARI, Cornell University

The ability to anticipate human actions is critical to many cyber-physical systems, such as robots and autonomous vehicles. Computer vision and sensing algorithms to date have focused on extracting and predicting visual features that are explicit in the scene, such as color, appearance, actions, positions, and velocities, using video and physical measurements, such as object depth and motion. Human actions, however, are intrinsically influenced and motivated by many implicit factors such as context, human roles and interactions, past experience, and inner goals or intentions. For example, in a sport team, the team strategy, player role, and dynamic circumstances driven by the behavior of the opponents, all influence the actions of each player. This article proposes a holistic framework for incorporating visual features, as well as hidden information, such as social roles, and domain knowledge. The approach, relying on a novel dynamic Markov random field (DMRF) model, infers the instantaneous team strategy and, subsequently, the players' roles that are temporally evolving throughout the game. The results from the DMRF inference stage are then integrated with instantaneous visual features, such as individual actions and position, in order to perform holistic action anticipation using a multi-layer perceptron (MLP). The approach is demonstrated on the team sport of volleyball, by first training the DMRF and MLP offline with past videos, and, then, by applying them to new volleyball videos online. These results show that the method is able to infer the players' roles with an average accuracy of 86.99%, and anticipate future actions over a sequence of up to 46 frames with an average accuracy of 80.50%. Additionally, the method predicts the onset and duration of each action achieving a mean relative error of 14.57% and 15.67%, respectively.

CCS Concepts: • **Computing methodologies** → **Machine learning approaches**; **Activity recognition and understanding;**

Additional Key Words and Phrases: Human, teams, action anticipation, computer vision, role inference, Markov random field, multi-layer perceptron, sports, hidden variables, domain knowledge

ACM Reference format:
Junyi Dong, Qingze Huo, and Silvia Ferrari. 2022. A Holistic Approach for Role Inference and Action Anticipation in Human Teams. *ACM Trans. Intell. Syst. Technol.* 13, 6, Article 95 (September 2022), 24 pages.
https://doi.org/10.1145/3531230

1 INTRODUCTION

As pointed out in the seminal work on mental cognition by Kenneth Craik in 1943 [15], animals utilize internal models of their external reality and of possible actions at their disposal in order

This work was supported by the Office of Naval Research under Grant No. N00014-17-1-2175.
Authors' address: J. Dong, Q. Huo, and S. Ferrari (corresponding author), Cornell University, 124 Hoy Road, Ithaca, New York 14850; emails: {jd979, qh223, ferrari}@cornell.com.

2157-6904/2022/09-ART95 $15.00
https://doi.org/10.1145/3531230

to evaluate various alternatives and conclude which one to utilize to react to new situations. In the context of teams and collaborative groups, individuals use their ability to anticipate human actions in a broad range of contexts and situations in order to decide their own subsequent actions and behaviors. Often, action anticipation is based on inferred cues, such as social roles, intentions, and goals that are deduced from visual information interpreted in the context of domain knowledge and past experiences. For example, people tend to choose their greetings, such as "shaking hands" or "hugging", based on their anticipation of the most likely response by the recipient [38]. Drivers routinely predict future actions of pedestrians, cyclists, and other drivers, based on their appearance, trajectories, driving style, and inferred social role, in order to guarantee safe driving [11, 12]. Similarly, athletes make split-second decisions based on the behavior of their teammates and opponents, their knowledge of the game, as well as their anticipation of opponents' actions [64]. As such, the ability to anticipate human actions is essential for human social life and bears great potential for future development of intelligent systems and machines. Team sports, in particular, provide an excellent benchmark problem for action anticipation because the rules and goals of the game are well-defined, video data is broadly available from event broadcasting, and players' decisions depend on many factors ranging from team strategy to individual roles, from knowledge of the game to opponent behaviors [47, 48].

In contrast to action recognition, which generates a semantic label from the video of an observed human behavior [18, 43, 54, 69], action anticipation aims at predicting one or more sequential human behaviors, several seconds into the future. Unlike traditional prediction algorithms, the approach presented in the article seeks to anticipate the semantic labels of a sequence of human actions before their onset, including sudden and radical behavioral changes such as switching from standing to hitting the ball. Existing methods for action anticipation can be categorized into feature-level, single-agent, and dual-agent anticipation. Feature-level anticipation predicts a convolutional feature representation of a future image for an ongoing action and, then, uses this representation to predict the action label classification [23, 49, 52, 55, 61]. These methods assume that a few initial frames of a human action is partially observed, based on which the remaining action sequences can be predicted. Moreover, feature-level anticipation relies primarily on prior data training and, therefore, fails in testing images that do not show globally similarity to the training data [62].

Single-agent anticipation predicts a semantic action label using appearance-based or motion-based features extracted from a sequence of frames preceding the onset of an action [9, 22, 50]. The input features can be enriched by incorporating information of the surrounding visual context, such as the presence of certain meaningful objects in the scene [39, 41]. A **long short-term memory** (**LSTM**) network was trained in [22] to predict an individual's cooking activity over the horizon of 0.25–2 s based on an observation time window of 1.75–3.5 s. The action anticipation performance of the cooking activity was quantitatively evaluated in [35] in terms of the observation duration and prediction horizon, showing that an increase in prediction horizon is accompanied by deterioration in anticipation accuracy even with long observations of up to 30 s.

Dual-agent action anticipation methods rely on extracting action-reaction patterns from videos of two-person interactions such as "hugging" or "pushing", in order to leverage the causal relationship in social interactions [6, 30, 38, 39]. However, the resulting algorithms are limited in scope in that the interaction is known *a priori*, and the anticipation is from the perspective of the reactive agent by only anticipating the reactive actions based purely on visual cues. The approach presented in this article is applicable to diverse forms of interactions among two or more persons, including team strategies and individual roles that evolve over time, and is capable of predicting action sequences and timing. Previous work has shown that the temporal localization of future events can be performed by learning a probability distribution of the occurrence time conditioned on a sequence of observed features [44]. In particular, the former method quantizes the prediction

horizon into discrete time intervals, one of which is predicted to contain the occurrence of the future event. One downside of such discrete-time model is the finite temporal resolution caused by quantization. As an improvement, a regression neural network was learned from data in [41, 44] to output a real positive value as a continuous approximate of the onset of the future action executed. In this article, the regression neural network is extended to the problem of predicting both the onset and duration of future actions in human teams.

Our holistic approach for interpreting and predicting team behaviors is demonstrated on a new and challenging problem, namely anticipating fast actions executed by interacting members of a sport team. In a team sport, such as volleyball, not only the team strategy and circumstances of play are hidden and directly influence individual actions, but also are highly dynamic, in that they change significantly and rapidly over time. Additionally, individual players assume different roles during the game, contributing in different measure to game strategy and outcome, thus influencing their teammates' behaviors in contrasting ways. The team strategy and players' roles are, almost by definition, hidden or unobservable. In other words, they are not visually explicit in the scene, but they can be inferred from a combination of visual cues and domain knowledge of the sport and of the team itself, as will be demonstrated in this article.

Inferring team strategy bears similarities to the problem of group activity recognition, which seeks to identify an activity label for a group of participants [31, 32, 56, 67]. However, these methods require the user to pre-select a time window that centers around a group activity by manually clipping the video or choosing the initial and final image frame. As such, they can not be easily extended to dynamic settings where the team strategies evolve over time, gradually or suddenly at unknown instants. In contrast, this article infers the team strategy label in each frame, based on which the input video can be automatically partitioned into scene segments for action anticipation.

On the other hand, role inference derives motivation from the "Role Theory" in sociology [40, 46, 58], which is a key concept for understanding the organization of social life and social activity. Recently, [25] defined roles as "socially defined expectations that a person in a given status follows", showing that roles provide predictability of people's behaviors. The importance of individual social roles in human events, such as "listener", "speaker", "bride", and "groom", has also been recently recognized in the computer vision literature [20, 46]. These methods, however, are not directly applicable to team action anticipation because they do not consider the rapid change in roles. Also, existing methods seek to label either the group activity or the individual role, whereas, in many events, such as sports, the individual role changes over time as a function of an evolving group activity/strategy. Furthermore, in many events, such as team sports, the interdependence between team strategies and players' roles cannot be necessarily categorized into a set of semantic classes identifiable *a priori*.

This article presents a novel **dynamic Markov random field** (**DMRF**) model that captures players' interrelationships using a dynamic graph structure, and learns individual player characteristics in the form of a feature vector based on a wealth of prior information, including domain knowledge, such as court dimensions and sport rules, and visual cues, such as homography transformations, and players' actions and jerseys. The DMRF unary and pairwise potentials can then be learned from data to represent the probability of individual feature realizations and the strengths of the corresponding players' interrelationships, respectively. Each new video frame is associated with a global hidden variable that describes the team strategy, within which each player is assigned a local hidden variable representing her/his role on the team. Then, given video frames of an ongoing game, the DMRF can be used to infer the players' roles using a **Markov chain Monte Carlo** (**MCMC**) sampling method, and to provide inputs to an **multi-layer perceptron (MLP)** that anticipates the players' future actions.

The notion of key player is introduced to distinguish a small set of players who will perform dominant actions that directly influence the game progress. In the anticipation stage, an MLP is trained to predict future actions of key players based on visual features as well as the inference results. Action anticipation is performed in each frame such that the anticipated results can be updated in a timely manner as the future unfolds. Inspired by recent work on predicting the temporal occurrence of future actions [41], the anticipation MLP is configured to simultaneously output the semantic label, onset and duration of the key players' future actions. In comparison to the existing research on single-agent and dual-agent action anticipation, this article raises a distinctively new variant of visual forecasting problem that anticipates future action in human teams. By proposing a new problem formulation and solution for team action anticipation, the holistic approach presented in this article allows to account for the implicit context, perceived through several inferred hidden variables, as well as for hybrid inputs comprised spatio-temporal relationships, continuous variables, and categorical features that together describe the team players and their interactions. The results obtained on testing database constructed from broadcasting videos of volleyball games demonstrate that this approach predicts the future actions of key players up to 46 frames into the future, with an accuracy of 80.50%. In addition, the approach achieves an average accuracy of 84.43% and 86.99% for inferring the team strategy and players' roles, respectively.

2 BACKGROUND AND PRELIMINARIES

The role inference and action anticipation approach presented in this article is demonstrated on the team sport of volleyball, described here briefly for completeness. However, the approach can be similarly applied to other team sports and activities, as will also be shown in future work. A volleyball match consists of five sets that are further broken into points. Each point starts with a player serving the ball to the opposite side. Each team must not let the ball be grounded within their own court by hitting the ball to the opponent after no more than three consecutive touches of the ball by three different players. The game continues until the ball is grounded, with the players moving around their own side of the court and assuming different roles over time, such as blocker, defense-libero, left-hitter, and so on (Figure 1). This alternating pattern can be reflected by the transition of a finite class of team strategy labels (Figure 1(a)), whose semantic meaning describes the technical activity of the two teams. For instance, the team strategy label in Figure 1(b) indicates that the right team is setting the ball for the next-step attack and the left team is on defense, whereas Figure 1(b) shows that the the right is attacking and the left is blocking.

The two teams are divided by a net in the middle of the court, which simplifies the action anticipation problem compared to other team sports, such as football or hockey, which will be studied in future work. Like other sports, each team is represented by a jersey color. But, in volleyball, some players within a team also wear a different jersey to indicate their "libero position" on the team. For effective coordination, players assume different roles in accordance to their expected duty in the team. Consequently, each player can be assigned a semantic role label that serves as an abstract representation of the player's intentions and possible actions. A complete description of the players' nine possible roles is shown in Figure 2. An important complexity is that the players roles change rapidly and unexpectedly over time, and some of the players can assume the same role at the same time.

Also volleyball actions can be categorized into nine well-defined classes: *spiking, blocking, setting, running, digging, standing, falling, waiting, and jumping*, extracted using computer vision algorithms [3, 4, 31, 32, 53]. However, actions are not unique to players' roles, nor there is any precise correspondence (e.g., one-to-one) between roles and actions. In this article, the action label *waiting* is replaced with *squatting* for a closer clarification on this defensive action that happens before a player digs the ball, as shown in Figure 3.

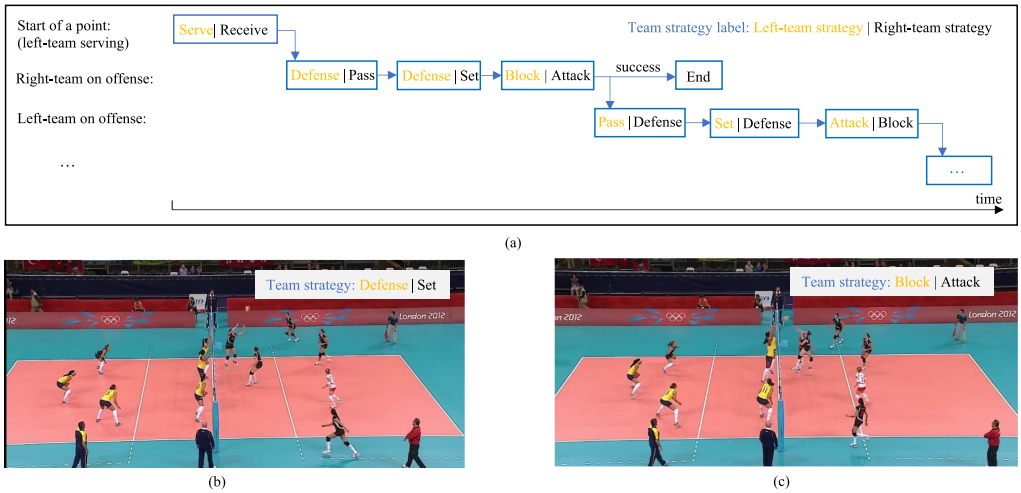


Fig. 1. Example of temporal evolution of team strategies in a volleyball match (a) and corresponding visual scenes (b–c).

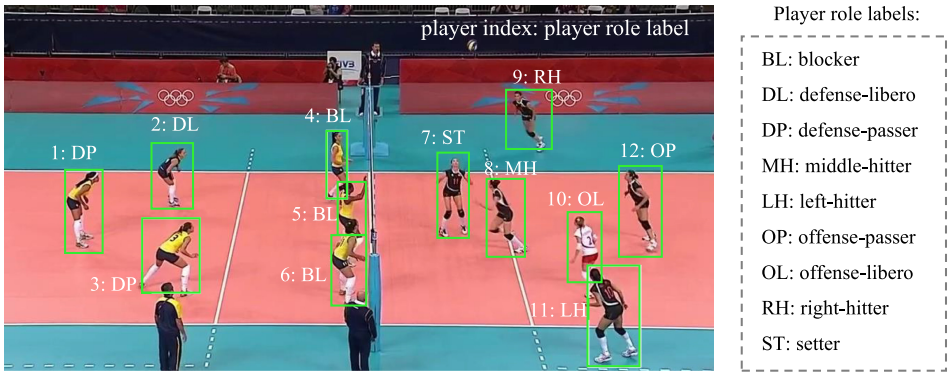


Fig. 2. Volleyball players' roles.

Fig. 3. Examples of nine volleyball players' actions.

During the volleyball match, players do not contribute equally. Rather, only a subset of players referred to as *key players* are actively engaged while the others are waiting for their turns to enter into action. For instance, player 7 in Figure 2 is a key player because her future action of *setting* will dominate the game.

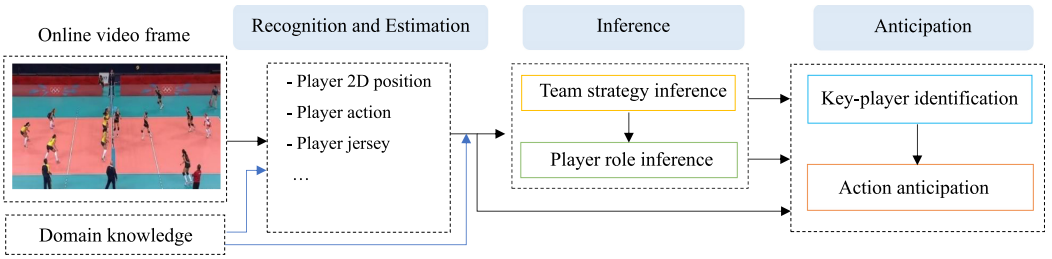


Fig. 4. A holistic framework for action anticipation in team sports.

3 PROBLEM FORMULATION AND ASSUMPTIONS

The problem addressed in this article consists of anticipating future actions by multiple key players in the team sport of volleyball based on hidden information, such as players' roles and team strategy, domain knowledge, and visual features extracted from video using existing computer vision algorithms [29, 31, 32, 36, 67, 68]. The goal is to develop a general and systematic approach for interpreting visual scenes of human group activities with complex goals, dynamic behaviors, and variegated interactions. Although this article mainly considers video data, the proposed framework can be readily applied to data obtained from other sensing modalities, such as range finders, inertial navigation units, and wearable sensors [28]. The approach is holistic in that it integrates image *recognition*, namely the classification of visually explicit information, state *estimation*, *inference* of hidden variables, and *anticipation* of future actions and events. As schematized in Figure 4, the approach consists of using the information extracted from domain knowledge (including prior videos) and streaming videos, using available image recognition and state estimation algorithms, to solve the problems of team/player inference and action anticipation problem formulated in Sections 3.1 and 3.2, respectively.

3.1 Inference Problem Formulation

Consider a video $\mathcal{V}$ comprised $K \in \mathbb{N}^+$ consecutive frames obtained at discrete moments with a constant sampling interval Δt. Each frame $I(k) \in \mathbb{R}^{h\times w}, k = 1, \ldots, K$, corresponds to an image matrix of $h \times w$ pixel intensities, where $h, w \in \mathbb{N}^+$ are the frame size. Let $\mathcal{N} = \{1, \ldots, N\}$, $N \in \mathbb{N}^+$, denote the index set of players extracted from $I(k)$ using computer vision [29, 32]. The frame index is omitted for $\mathcal{N}$ since the number of players is fixed in a volleyball video.

Each player in frame $I(k)$ can be associated with an index $i \in \mathcal{N}$ and a feature descriptor that contains a 2D position vector, an action label, and an appearance feature describing the player's jersey color. Other characteristics and state variables can be similarly included, depending on the application of interest. Let $\mathbf{p}'_i(k) = [x'_i(k) \quad y'_i(k)]^T \in \mathbb{R}^{2\times 1}$ denote the $2D$ position of the ith player with respect to the image frame, which can be approximated by the image coordinate at the bottom middle point of the player's bounding box. In order to gain immediate insight into players' spatial relationship, the position vector $\mathbf{p}'_i(k)$ is resolved into the inertial coordinate denoted by $\mathbf{p}_i(k) = [x_i(k) \quad y_i(k)]^T \in \mathbb{R}^{2\times 1}$. Because the volleyball court is planar, the image and inertial coordinate can be related via homograph transformation H, as shown in Figure 5,

$$\lambda \begin{bmatrix} x'_i(k) \\ y'_i(k) \\ 1 \end{bmatrix} = \begin{bmatrix} H_{11} & H_{12} & H_{13} \\ H_{21} & H_{22} & H_{23} \\ H_{31} & H_{32} & H_{33} \end{bmatrix} \begin{bmatrix} x_i(k) \\ y_i(k) \\ 1 \end{bmatrix}, \tag{1}$$

where $\lambda \neq 0$ is a scaling factor, and the homography matrix H can be estimated using domain knowledge of court dimensions and the geometry of the lines drawn on the volleyball court [19, 27, 59, 60].

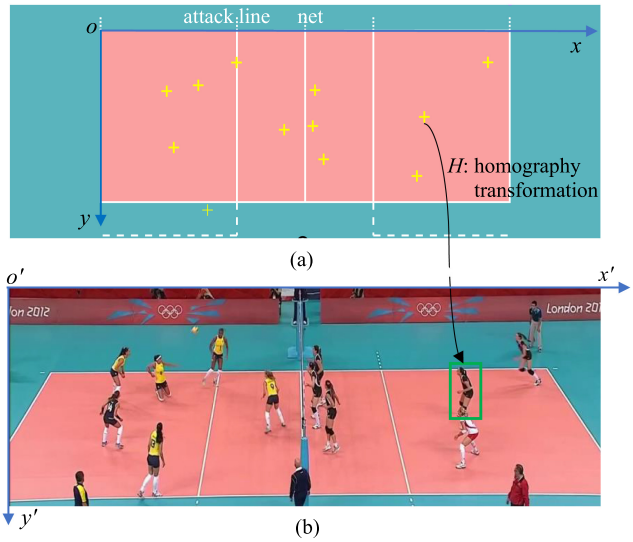


Fig. 5. Projection between the inertial reference frame (a) and image reference frame (b).

Next, let $A_i(k) \in \mathcal{A}$ represent the action label of player $i \in \mathcal{N}$ in an observed frame $I(k)$, where $\mathcal{A}$ is the discrete and finite range of the action classes shown in Figure 3. A player's jersey color is denoted by a discrete variable $C_i(k) \in C$, which can be obtained using a color detector [13, 33, 57] or as prior knowledge. Together, the aforementioned features can be organized as a player feature vector $F_i(k) = [\mathbf{p}_i(k)^T \quad A_i(k) \quad C_i(k)]^T$.

Then, each frame $I(k) \in \mathcal{V}$ in a volleyball video can be assigned a semantic label describing the technical strategy of two teams, as illustrated in Figure 1(b–c). Inference of the team strategy requires the aggregation of features across players, which amounts to the concatenation of player feature vectors into a frame-wise team descriptor. In order to preserve the spatial relationship in a team, feature vectors of players on each side are sorted by the player's distance to the net. Then, the aggregated team feature descriptor can be constructed as

$$F(k) \triangleq [F_{l_1}^T(k) \quad \dots \quad F_{l_{\frac{N}{2}}}^T(k) \quad F_{r_1}^T(k) \quad \dots \quad F_{r_{\frac{N}{2}}}^T(k)]^T \tag{2}$$

with the range denoted by $\mathcal{F}$ and the indices of elements defined by the sorted index set

$$\hat{\mathcal{N}} = \{l_1, \dots, l_{\frac{N}{2}}, r_1, \dots, r_{\frac{N}{2}}\}, \tag{3}$$

where $\{l_1, \dots, l_{\frac{N}{2}}\} \subset \hat{\mathcal{N}}$ represent the sorted indices of players on the left team and $\{r_1, \dots, r_{\frac{N}{2}}\} \subset \hat{\mathcal{N}}$ is the counterpart for the right team.

Let $S(k) \in \mathcal{S}$ be a global hidden variable representing the team strategy label in frame $I(k)$, where $\mathcal{S}$ is the finite range of the team strategy classes, as illustrated in Figure 1. In addition, let $X_i(k) \in \mathcal{R}, i \in \mathcal{N}$, be a local hidden variable representing the role of player i. $X_i(k)$ takes a realization from a set of role labels $\mathcal{R}$, which are illustrated in Figure 2. The labels of all players' roles can be denoted by a random vector $X(k) \triangleq [X_1(k) \quad \dots \quad X_N(k)]^T$ that has range $\mathcal{X} = \mathcal{R}^N$. Then, the inference problem can be formulated as follows:

Problem 1. *Given the extracted features, $F(k)$, learn a multi-class classifier, $f_S : \mathcal{F} \to \mathcal{S}$, that maps $F(k) \in \mathcal{F}$ to a team strategy label $S(k) \in \mathcal{S}$. Subsequently, learn an inference model,*

Table 1. Notation of Frame Variables and Segment Variables

Frame variable	Description	Segment variable	Description
$S(k)$	Team strategy in frame $I(k)$	$S_l = \{S(k) \mid k \in T_l\}$	Team strategy in segment V_l
$A_i(k)$	Action of player i in frame $I(k)$	$A_{i,l} = \{A_i(k) \mid k \in T_l\}$	Action of player i in segment V_l
$X_i(k)$	Role of player i in frame $I(k)$	$X_{i,l} = \{X_i(k) \mid k \in T_l\}$	Role of player i in segment V_l
$\mathbf{p}_i(k)$	2D location of player i in frame $I(k)$	$P_{i,l} = \{\mathbf{p}_i(k) \mid k \in T_l\}$	2D location of player i in segment V_l

$f_X : \mathcal{F} \times \mathcal{S} \rightarrow \mathcal{X}$, that maps the feature vector $F(k)$ and the inferred team strategy label $S(k)$ to the vector $X(k)$, representing role labels of all players.

3.2 Anticipation Problem Formulation

The goal of the action anticipation problem is to leverage the confluence of information including inferred team strategies, inferred players' roles and features, as well as domain knowledge, in order to predict which are the key players and what are their respective future action sequences. Given the inferred team strategy up to the current frame, κ, (obtained from *problem 1*), a scene change point is defined as a frame index τ such that

$$S(\tau) \neq S(\tau + 1), \quad \tau = 1, \ldots, \kappa - 1 \tag{4}$$

and is typically unknown *a priori*. Let $\boldsymbol{\tau} = [\tau_1 \ldots \tau_m]T$ represent the scene change points up to the current time κ, where $\tau_1 = 1$ and $\tau_m \leq \kappa$. Video frames between every two consecutive scene change points have the same inferred team strategy and, therefore, can be automatically grouped as a scene segment, which eliminates the algorithm's dependence on pre-trimmed videos. Let $V_l, l = 1, \ldots, m$ denote the lth scene segment with the frame-index set T_l defined as

$$T_l = \begin{cases} \{\tau_l, \ldots, \tau_{l+1} - 1\} & l = 1, \ldots, m-1 \\ \{\tau_l, \ldots, \kappa\} & l = m \end{cases} \tag{5}$$

Consequently, V_l can be represented as

$$V_l = \{I(k) \mid k \in T_l\}, \quad l = 1, \ldots, m. \tag{6}$$

The duration of V_l, denoted by d_l, equals the number of frames in T_l multiplied by the discrete-time sampling interval Δt

$$d_l = \begin{cases} (\tau_{l+1} - \tau_l)\Delta t & l = 1, \ldots, m-1 \\ (\kappa - \tau_l + 1)\Delta t & l = m \end{cases}. \tag{7}$$

After defining the scene segments, variables that are defined in each frame $I(k)$ can be upgraded to represent the whole segment, as shown in Table 1, where the argument in "()" represents the frame index, the subscript "i" represents the player index, and the subscript "l" represents the segment index.

In order to distinguish a small set of players who will perform dominant actions that influence the game progress, a binary indicator variable $\mu_i(\kappa) \in \{0, 1\}$ is introduced for a player i such that its value equals one if the corresponding player will become a key player, and equals zero otherwise. $\mu_i(\kappa)$ can be obtained by constructing a mapping, $f_\mu : \mathcal{S} \times \mathcal{R} \rightarrow \{0, 1\}$, that takes as input the inferred team strategy label $S(k)$ and role label $X_i(k)$ and outputs the binary indicator value

$$\mu_i(\kappa) = f_\mu(S(k), X_i(k)) \tag{8}$$

$f_\mu(\cdot)$ can be learned as a binary classifier based on a small amount of annotated data, or it can be derived using domain knowledge about the likelihood of a player being the key player given the corresponding role and team strategy. The complete set of predicted key players is

$$\mathcal{K} = \{i \mid \mu_i(\kappa) = 1,\ i \in \mathcal{N}\}. \tag{9}$$

Action anticipation of a key player considers four types of information collected in the current scene segment V_m, i.e., the inferred team strategy S_m, the inferred role $X_{i,m}$, the ongoing action $A_{i,m}$ and the player's 2D spatial location $P_{i,m}$. Furthermore, the Markov assumption is adopted such that future action $A_{i,m+1}$, is independent from the past action $A_{i,m-1}$ with given $\{A_{i,m}, P_{i,m}, X_{i,m}, S_m\}, i \in \mathcal{K}$. The Markov assumption is justifiable because the hybrid inputs encode information from multiple sources, hence enriching the model and reducing the dependence of future action on historical data. By virtue of such assumption, action anticipation only requires a short-term input with arbitrary starting scenes. Finally, the action anticipation problem can be summarized as follows:

Problem 2. *Given the inferred team strategy label $S(\kappa)$ and role label $X(\kappa)$ of the current frame $I(\kappa) \in V_m$, predict the set of key players, $\mathcal{K} \subseteq \mathcal{N}$, using (8–9). Then, for each key player $i \in \mathcal{K}$, predict the semantic label, onset and duration of their future actions $A_{i,m+1}$ using aggregated input sequences $\{A_{i,m}, P_{i,m}, X_{i,m}, S_m\}$.*

4 INFERENCE MODEL

Inferring team strategy requires a multi-class classifier to map the feature vector $F(k)$ to a label $S(k)$ that represents the technical team activity in each frame. This article uses an MLP to perform the task while other classifiers such as random forests [63] are also applicable. The inferred team strategy label, $S(k)$, is appended to the feature vector of the ith player to form an augmented feature vector, i.e., $Z_i(k) = [F_i(k)^T \quad S(k)]^T, i \in \mathcal{N}$, which can then be organized into an augmented feature matrix for all players

$$Z(k) = [Z_i(k) \quad \dots \quad Z_N(k)]. \tag{10}$$

This section develops a novel DMRF model with dynamical graph structures for inferring the joint probability of players' roles $X(k)$ from the augmented feature matrix $Z(k)$.

4.1 Dynamic Markov Random Field (DMRF) Model of Team Player Roles and Interactions

Classic MRFs are probabilistic models comprised an undirected graph with a set of nodes that each represent correlated random variables, and a set of undirected arcs (i.e., graph structure) that represent a factorization of the joint MRF probability learned from data [21]. The advantages of MRFs over other probabilistic models are that they can model processes with both hidden and observable variables, as well as include both categorical and continuous variables by describing different types of relationships using unary and pairwise potentials. MRF was introduced into the image processing field in the 1980s [24] and was henceforth widely used in computer vision problems such as image segmentation [26, 45], image denoising [8], and image reconstruction [10, 42]. While in classic MRFs, the graph structure is fixed and decided *a priori*, this article presents an approach for constructing dynamic MRFs (or DMRFs) representations of the visual scene. The goal is to learn a temporally evolving graph structure from each frame for the inference of hidden role variables, where only the set of nodes remains unchanged, and the arcs appear or disappear from frame to frame based on the events in the scene.

In this approach, every hidden node, denoted by $X_i(k)$ ($i \in \mathcal{N}$), represents the hidden role of player i, and every observable node, denoted by $Z_i(k)$ ($i \in \mathcal{N}$), represents the feature vector of player i. The temporally evolving arc set, $\mathcal{E}(k)$, is then learned from the players' relative distance by minimizing an energy function such that the minimum value corresponds to the optimal arc configuration. In order to infer the players' roles from all available information, each node $X_i(k)$ is connected to the corresponding feature vector $Z_i(k)$. $X_i(k)$ is associated with a unary potential $\phi(X_i(k), Z_i(k))$ that captures how probable feature $Z_i(k)$ is for different realizations of $X_i(k)$.

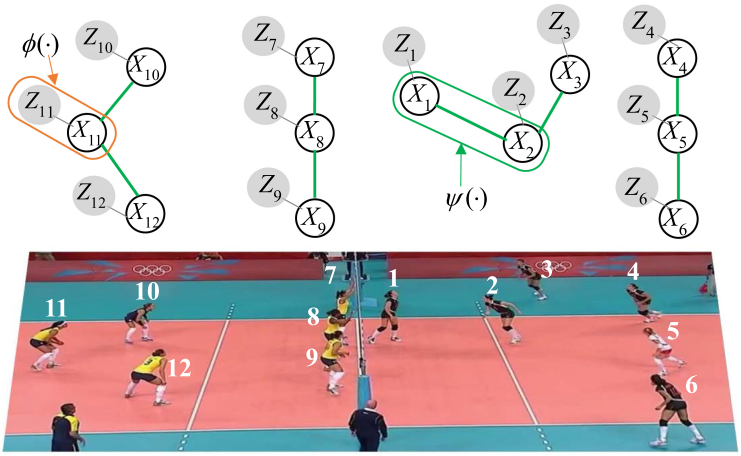


Fig. 6. DMRF model for player role inference, where the time argument k is omitted for brevity.

Every arc is associated with a pairwise potential $\psi(X_i(k), X_j(k))$ that represents the strength of correlations between the two random variables $X_i(k)$ and $X_j(k)$ in a spatial neighborhood. Then, the joint probability distribution of the random variables can be factorized as the product of potential functions over the graph structure [37, 66]

$$P(X(k)|Z(k), \mathcal{E}(k)) = \frac{1}{C} \prod_{i \in \mathcal{N}} \phi(X_i(k), Z_i(k)) \prod_{i,j \in \mathcal{E}(k)} \psi(X_i(k), X_j(k)), \tag{11}$$

where C is the partition function that guarantees $P(X(k)|Z(k))$ is a valid distribution and the scope of pairwise potentials is determined by the estimated graph structure $\mathcal{E}(k)$. An example of DMRF graph representation is illustrated in Figure 6 and the potential functions are learned as explained in the following subsections.

4.1.1 DMRF Potential Functions. The unary potential $\phi(X_i(k), Z_i(k))$ expresses how probable the feature vector $Z_i(k)$ is for different realization of the role label $X_i(k)$, and can be modeled as a likelihood function [5, 37, 51],

$$\phi_i(X_i(k), Z_i(k)) \triangleq P(Z_i(k)|X_i(k)). \tag{12}$$

Let $\mathcal{R} = \{1, 2, \ldots, R\}$ denote the set of role labels such that $X_i(k) = n$ ($n \in \mathcal{R}$) if player i assumes the nth semantic role label. Let $\mathbf{1}_n \in \{0, 1\}^R$ be a R-dimensional one-hot vector where the nth entry equals one and the rest entries equal zero. The likelihood function can be defined as

$$P(Z_i(k)|X_i(k) = n) = \frac{\exp\{\mathbf{1}_n^T \cdot [W_{u2} \cdot \sigma(W_{u1} \cdot Z_i(k))]\}}{\sum_{m=1}^{R} \exp\{\mathbf{1}_m^T \cdot [W_{u2} \cdot \sigma(W_{u1} \cdot Z_i(k))]\}}, \tag{13}$$

where $\sigma(\cdot)$ is the sigmoid function, W_{u1} and W_{u2} are weights that will be learned from data and their dimensions are hyper-parameters selected to agree with the dot product.

Pairwise potential concerns the interrelationship between two node variables taking particular roles, with greater value indicating higher probability for the corresponding players to interact in a team. For instance, the pair "setter - hitter" has a higher chance to interact in a close proximity than "setter - blocker" pair since the latter only appears in two opposing teams. Let $W_p \in \mathbb{R}^{R \times R}$ denote the weight matrix that represents the correlation between a pair of roles. Then, the pairwise

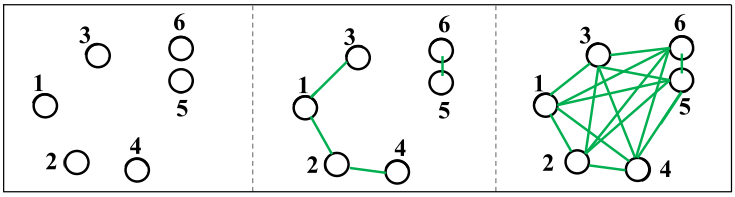


Fig. 7. A graphical model of six nodes with an empty arc set (a), a sparse arc set (b), and a dense FC arc set (c).

potential is defined as

$$\psi(X_i(k) = n, X_j(k) = m) \triangleq \mathbf{1}_n^T \cdot W_p \cdot \mathbf{1}_m. \tag{14}$$

4.1.2 DMRF Graph Structure. The graph structure, $\mathcal{E}(k)$, determines the scope of pairwise potentials. Traditionally, the MRF graph structure is established a *priori* and remains fixed (e.g., [65, 66]). In order to use MRF models for dynamic role inference, a new approach is developed here to learn and adapt the structure online based on streaming video frames. In this approach, the structure can vary from an empty arc set to a **fully connected** (**FC**) configuration, as shown in Figure 7. An empty arc set (Figure 7(a)) indicates that all nodes (e.g., players' roles) are independent and there are no interactions between them. Conversely, a densely connected configuration (such as that in Figure 7(c)) captures many interrelationships, including redundant ones and, thus, may incur unnecessary computational burden. The approach developed in this article produces an efficient structure estimation algorithm (16–20) to dynamically estimate a sparse structure (Figure 7(b)) that captures only the most significant interactions in each video frame.

Let $Y_{i,j}(k)$ denote a binary variable such that its value $y_{i,j}(k)$ equals one when an interaction arc exists between players labeled by i and j, and equals zero otherwise. Then the arc set can be denoted as $\mathcal{E}(k) = \{(i,j)|y_{i,j}(k) = 1, i,j \in \mathcal{N}\}$, and the structure estimation problem can be cast as a constrained optimization problem over the arc variables $Y_{i,j}(k)$. In many human team activities, such as sports, proximity is an indication of potential interactions and, therefore, in this article the DMRF graph structure is indicative of interrelationships between spatial neighbors. Other representations are also possible, depending on the application, and may be adopted in the proposed approach with small modifications. Then, the Euclidean distance $d_{i,j}(k) = \|\mathbf{p}_i(k) - \mathbf{p}_j(k)\|$ between every pair of players is used to construct an energy function that is linear in the realizations of the arc variables $Y_{i,j}(k)$,

$$E(Z(k), \mathcal{E}(k)) \triangleq \sum_{(i,j)\in\mathcal{E}(k)} d_{i,j}(k)\, y_{i,j}(k) \tag{15}$$

such that the optimal arc configuration corresponds to the minimum of the energy function. Subsequently, minimizing the energy function can be approached by solving an Integer Linear Program

$$\min_{\mathcal{E}(k)} \quad \sum_{(i,j)\in\mathcal{E}(k)} d_{i,j}(k) y_{i,j}(k) \tag{16}$$

$$y_{i,j}(k) = y_{j,i}(k), \quad \forall (i,j) \in \mathcal{E}(k) \tag{17}$$

$$\text{sbj to} \quad \sum_{i\in\mathcal{N}} y_{i,j}(k) \geq 1, \quad \forall j \in \mathcal{N} \tag{18}$$

$$\sum_{i \in \mathcal{N}} y_{i,j}(k) \leq 2, \qquad \forall j \in \mathcal{N} \tag{19}$$

$$y_{i,j}(k) \in \{0,1\}, \qquad \forall (i,j) \in \mathcal{E}(k). \tag{20}$$

The constraint in (17) guarantees that interactions are symmetric, and (18)–(19) specify that a node has a minimum of one and maximum of two arcs connecting to its spatial neighbours, resulting in a sparse structure. Although only the proximity feature is considered, the proposed method is a generic algorithm that can incorporate other features to estimate social interactions. Details are referred to the previous work [17]. After $\mathcal{E}(k)$ is estimated, the joint probability distribution of the role variables in (11) is factorized as the product of potential functions over $\mathcal{E}(k)$.

4.2 Spatio-temporal MRF Model

In this subsection, an approach is presented for reconstructing the temporal evolution of random variables $X(k)$ across frames to recursively estimate the joint role labeling using a sequence of feature vectors and the DMRF model of a single frame derived in (11). Let $\gamma(X_i(k-1), X_i(k))$ denote the temporal potential function that measures the compatibility of temporal transitions between $X_i(k-1)$ and $X_i(k)$. The temporal potential function can be modeled by a transition matrix $W_t \in \mathbb{R}^{R \times R}$ such that

$$\gamma(X_i(k-1) = n, X_i(k) = m) \triangleq \mathbf{1}_n^T \cdot W_t \cdot \mathbf{1}_m. \tag{21}$$

The temporal potential function can be integrated with the pairwise potential function to construct a joint state transition function

$$P(X(k)|X(k-1)) \propto \prod_{i \in \mathcal{N}} \gamma(X_i(k-1), X_i(k)) \prod_{i,j \in \mathcal{E}(k)} \psi(X_i(k), X_j(k)). \tag{22}$$

On the other hand, the product of unary potentials can be treated as the joint likelihood function, assuming that individual features are conditionally independent given the realization of random variables

$$P(Z(k)|X(k)) = \prod_{i \in \mathcal{N}} P(Z_i(k)|X_i(k)) = \prod_{i \in \mathcal{N}} \phi(X_i(k), Z_i(k)). \tag{23}$$

Let $Z(1,k) = \{Z(l)|1 \leq l \leq k\}$ denote a sequence of extracted feature vectors obtained from an initial frame ($l = 1$) up to the kth frame. Then, the joint probability of $X(k)$ can be recursively estimated from $Z(1,k)$ in a fashion similar to Bayesian filtering [16]

$$P(X(k)|Z(1,k)) = \frac{1}{\hat{C}} P(Z(k)|X(k)) \sum_{X(k-1)} P(X(k)|X(k-1)) P(X(k-1)|Z(1,k-1)), \tag{24}$$

where $\hat{C}$ is the partition function that guarantees $P(X(k)|Z(1,k))$ is a valid distribution. The proposed spatio-temporal MRF model is illustrated in Figure 8. The challenge arises because $P(X(k)|Z(1,k))$ is a multi-dimensional joint distribution that has significant computational ramifications. In order to keep the computation tractable, the joint distribution is achieved via the MCMC sampling method [1, 7, 14] by constructing a set of random samples that constitute a Markov chain whose stationary distribution converges to the desired distribution.

4.3 Learning of Potential Functions

The MRF model is trained in an incremental manner [2] in which the parameters of unary potentials are first trained and then fixed to learn the pairwise potentials. This incremental training allows the pairwise potentials to be built upon strong unary potentials, which makes the training

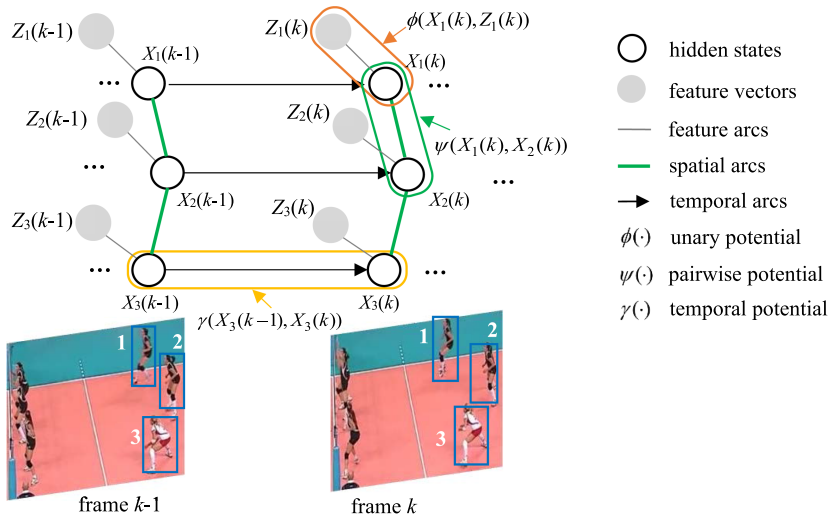


Fig. 8. Spatio-temporal MRF model for modeling players' roles.

more efficient because otherwise the pairwise potentials may not be able to capture the significant interactions from misleading unary potentials. In particular, the unary potential is trained by minimizing the cross entropy loss function, whereas the pairwise potential can be learned using the structural support vector machine framework [17, 34] or using domain knowledge about the relationship between different roles. This two-stage learning is performed in a frame-wise manner by leaving out the temporal transition matrix, which is fine-tuned at last on the training database. This incremental training allows the model to learn specific information presented in each potential function [2] and reduces the computational burden that would otherwise be incurred if all potential functions are learned together.

4.4 MCMC Inference

Inferring a role labeling $X(k)$ from the joint distribution $P(X(k)|Z(1,k))$ suffers from an enormous combinatorial complexity. Naively searching through the set of all possible labeling is intractable because the set has a cardinality that is exponential in the number of states. This article adopts the MCMC method [1, 14] to address the computational ramifications, which generates a Markov chain over the space of the joint configuration $X(k)$, such that the chain has a stationary distribution converging to $P(X(k)|Z(1,k))$. Assume the posterior $P(X(k-1)|Z(1,k-1))$ at time $k-1$ is represented by a set of $N_s \in \mathbb{R}^+$ samples $\{X(k-1)^{(\ell)}\}_{l=1}^{N_s}$, and each sample corresponds to a joint role labeling of all players, i.e., $X(k-1)^{(\ell)} = [X_1(k-1)^{(\ell)} \quad \dots \quad X_N(k-1)^{(\ell)}]^T$. Then, the Monte Carlo approximation to the posterior distribution in (24) at time k is

$$P(X(k)|Z(1,k)) \approx \frac{1}{\hat{C}} P(Z(k)|X(k)) \sum_{\ell=1}^{N_s} P(X(k)|X(k-1)^{(\ell)}). \tag{25}$$

Substitute (22–23) into (25), which gives

$$P(X(k)|Z(1,k)) \approx \frac{1}{\hat{C}} \prod_{i \in \mathcal{N}} \phi(X_i(k), Z_i(k)) \prod_{i,j \in \mathcal{E}(k)} \psi(X_i(k), X_j(k)) \sum_{\ell}^{N_s} \prod_{i} \gamma(X_i(k-1)^{(\ell)}, X_i(k)) \tag{26}$$

resulting in a sample-based representation for the distribution $P(X(k)|Z(1,k)) \approx \{X(k)^{(\ell)}\}_{\ell=1}^{N_s}$. The **Metropolis-Hastings** (**MH**) algorithm with the symmetric random walk proposal distribution [1, 14] is implemented for simulating the Markov chain.

5 ANTICIPATION MODEL

The goal of action anticipation is to predict a set of key players and their future actions as time evolves. Existing methods can not be easily adapted to the action anticipation problem (*problem 2*) because they do not take into account the time varying team strategy and players' roles, which are core to team actions. The anticipation model presented in this article differs from the existing methods by the input information exploited, which aggregates inferred hidden variables (inferred team strategy and players' roles) with explicit visual features, forming a rich input representation. The prediction of key players, $\mathcal{K} \subset \mathcal{N}$, is first achieved via (8–9). Subsequently, for each predicted key player, $i \in \mathcal{K}$, the action anticipation model merges four types of information corresponding to the current scene segment, i.e., $\{S_m, X_{i,m}, A_{i,m}, P_{i,m}\}$, to anticipate the future action $A_{i,m+1}$. The representation of input segments directly affects the learning efficiency and computational cost of the model. Thus, it is worth exploring a compact representation of $\{S_m, X_{i,m}, A_{i,m}, P_{i,m}\}$. Based on the definition of the scene change point and scene segment in (4–6), the segment variable of team strategy, S_m (Table 1), takes a constant value within the scene segment V_m. Hence, S_m can be fully defined by its value at the current time, κ, and the duration of V_m up to κ, that is, $S_m \triangleq (S(\kappa), d_m)$. Although values of $X_{i,m}$, $A_{i,m}$, and $P_{i,m}$ can vary within a scene segment, it is observed that future actions are most closely related to their respective values at the current time κ. Furthermore, this article seeks a frame-wise representation of the anticipation input and output, such that they can be updated instantaneously as time unfolds. As a result, only $A_i(\kappa)$, $X_i(\kappa)$, and $\mathbf{p}_i(\kappa)$ are preserved as inputs, as shown in Figure 9, which, together with $(S(\kappa), d_m)$, constitute an input vector

$$\mathbf{u}_i(\kappa) = [S(\kappa) \quad X_i(\kappa) \quad A_i(\kappa) \quad \mathbf{p}_i(\kappa)^T \quad d_m]^T, \tag{27}$$

where the time-varying characteristic of d_m represents the variable duration of the team strategy $S(\kappa)$. Likewise, the anticipation output, $A_{i,m+1}$, is designed to have an instantaneous representation of the future actions. Let t_s denote the *time to onset*, that is, the amount of time until the onset of $A_{i,m+1}$, and let d_{m+1} denote the duration of $A_{i,m+1}$. Then, $A_{i,m+1}$ can be defined as $A_{i,m+1} \triangleq (A_i(\kappa + t_s), d_{m+1})$, as shown in Figure 9(b). Equivalently, $A_{i,m+1}$ can be specified by a vector representation comprising three unknown variables

$$\mathbf{y}_i(\kappa) = [A_i(\kappa + t_s) \quad t_s \quad d_{m+1}]^T \tag{28}$$

It follows from (27–28) that the goal of the action anticipation task is to predict $\mathbf{y}_i(\kappa)$ based on $\mathbf{u}_i(\kappa)$ as time evolves.

An MLP is designed to perform the anticipation task based on the proposed input-output representation in (27–28). Categorical variables in $\mathbf{u}_i(\kappa)$ are converted to binary representations via one-hot encoding. The encoded $\mathbf{u}_i(\kappa)$ is passed through two branches, as shown in Figure 10, where the top branch is configured to output a probabilitys distribution for the discrete variable $A_i(\kappa + t_s)$ and the bottom branch generates two positive scalar values for the continuous variables, t_s and d_{m+1}, respectively. In particular, the top branch first maps the input vector to a latent vector, $\mathbf{h}_1$, using a FC layer followed by the relu-activation function

$$\mathbf{h}_1 = relu(W_{h1}\mathbf{u}_i(\kappa)), \tag{29}$$

where W_{h1} is the weight matrix. Subsequently, $\mathbf{h}_1$ is fed to the output layer, composed of a FC layer and the softmax activation function, to generate the conditional probability distribution of

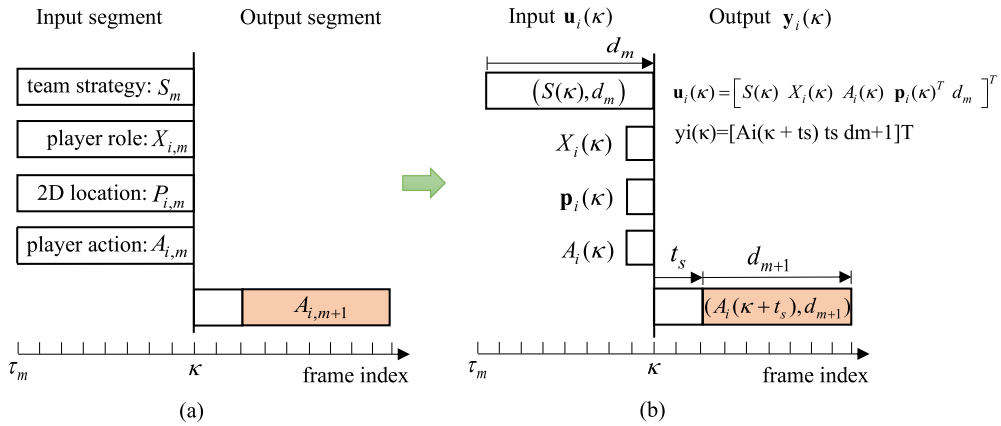


Fig. 9. Input and output segment for action anticipation of the ith key player (a) and the simplified instantaneous representation (b).

$P(A_i(\kappa+t_s)|\mathbf{u}_i(\kappa))$. Let $\mathcal{A} = \{1, 2, \ldots, A\}$ denote the range of the action classes, where each integer, $a \in \mathcal{A}$, represents a semantic action label, and $W_{o1} = [\mathbf{w}_1 \quad \ldots \quad \mathbf{w}_A]^T$ denote the weight matrix of the output FC layer. Then, $P(A_i(\kappa + t_s) = a|\mathbf{u}_i(\kappa))$ is computed as

$$P(A_i(k + t_s) = a|\mathbf{u}_i(k)) = \frac{\exp(\mathbf{w}_a^T \mathbf{h}_1)}{\sum_{a'=1}^{A} \exp(\mathbf{w}_{a'}^T \mathbf{h}_1)}, \quad a \in \mathcal{A} \tag{30}$$

and the action class with the highest probability is chosen as the anticipated action. Although the bottom branch adopts the same structure as the top branch, the FC-layers can have different dimensions and the output activation function is designed to be a relu-activation function for guaranteeing real positive values of t_s and d_{m+1}. Let W_{h2} denote the weights of the hidden FC layer in the bottom branch, and $W_{o2} = [\mathbf{w}_\tau \quad \mathbf{w}_d]^T$ denote the weights of the corresponding output FC layer. Then, t_s and d_{m+1} are obtained as follows:

$$\begin{aligned} \mathbf{h}_2 &= relu\ (W_{h2}\mathbf{u}_i(\kappa)). \\ t_s &= relu\ (\mathbf{w}_\tau^T \mathbf{h}_2) \\ d_{m+1} &= relu\ (\mathbf{w}_d^T \mathbf{h}_2) \end{aligned} \tag{31}$$

The complete set of the MLP parameters, $\Theta_A = \{W_{h1}, W_{h2}, W_{o1}, W_{o2}\}$, is trained by minimizing an anticipation loss that is a function of the ground truth and the actual predicted output. In particular, the loss function is formulated as the summation of the cross-entropy loss of the discrete action variable, $A_i(\kappa + t_s)$, and the mean squared loss of the two timing variables, t_s and d_{m+1}.

In summary, the input-output representation in (27–28) allows the input to be updated in each frame and the anticipation output to progressively change as more observations stream in. Furthermore, the trained model is shared across all players, and, therefore, anticipation for multiple players can be performed simultaneously by constructing an input vector for each of them.

6 EXPERIMENTS

In this section, experiments are conducted in order to validate the accuracy of the proposed methods. Using the Volleyball Activity Dataset [32], a supervised training database for the proposed inference and anticipation algorithms was obtained by annotating team strategies, player's roles,

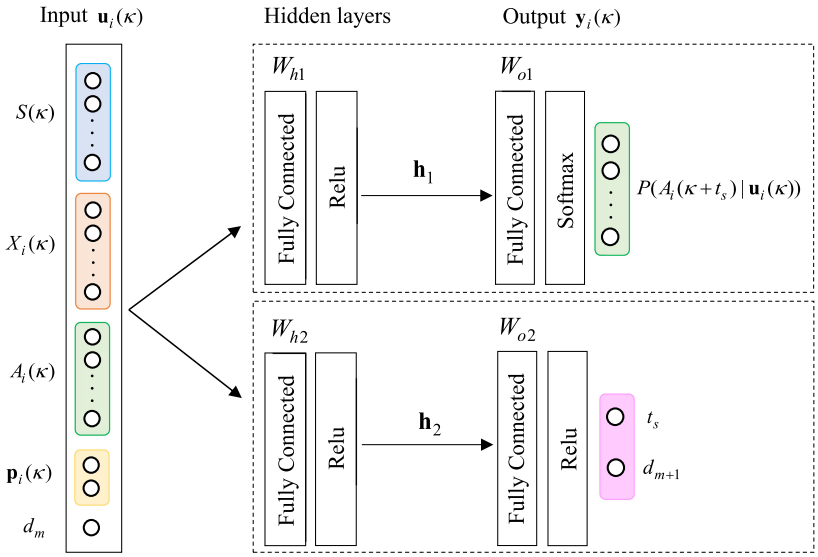
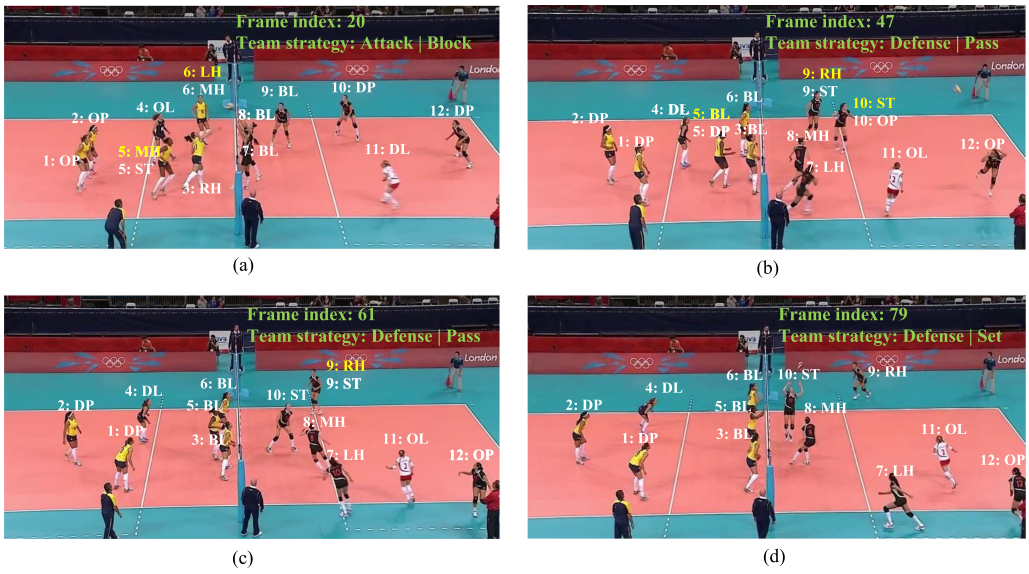


Fig. 10. MLP for action anticipation.

player's actions, and other necessary visual and positional information. Despite additional supervision required for learning the intermediate hidden variables, the overall labeling effort is less than that required by deep neural network models for action anticipation trained solely on images. The reason is that the proposed approach exploits the problem structure and incorporates domain knowledge before training the DMRF and MLP models. The inference and anticipation results are analyzed qualitatively and quantitatively on the testing data. Comparison with existing work on action anticipation was unfortunately not possible because existing algorithms are only applicable to single-agent or dual-agent activities [9, 22, 41, 50]. Therefore, the experiments in this article focused on evaluating the overall performance of the inference and anticipation model. Moreover, comparative studies (Section 6.2) that involve three types of experiments are carried out to determine the anticipation performance variability as a function of the hidden variables and corresponding inference accuracy.

6.1 Inference and Action Anticipation Results

The DMRF inference results are shown in Figure 11 for a sample sequence of frames extracted from a testing video clip, where the inferred team strategies and players' roles evolve over time. Notice that a team strategy spans over several consecutive frames, during which the action and spatial layout of players may be shifted, but not qualified to be inferred as a different category. The DMRF model presented in Section 4 correctly infers that the team strategy changes from "attack | block" (Figure 11(a)) to "defense | pass" (Figure 11(b–c)) to "defense | set" (Figure 11(d)), exemplifying the algorithm's robustness to the dynamically evolving scenes. Similarly, the players' roles change as the game unfolds. For example, the role of player 3 alters from "right-hitter" to "blocker", whereas player 7, originally a "blocker", becomes a "left-hitter". For comparison purposes, ground truth labels of the false inference results are shown in yellow above the (white) inferred roles in Figure 11. It is seen that inference failures are likely to happen when players are shifting to new locations. For instance, the algorithm mistakenly infers the roles of player 9 and 10 in Figure 11(b). However, as more observations are received, the updated inference results would be self-corrected and thus



BL: blocker, DP: defense-passer, DL: defense-libero, MH: middle-hitter, LH: left-hitter, OP: offense-passer, OL: offense-libero, RH: right-hitter, ST: setter

Fig. 11. Evolution of the inferred team strategy from "attack | block" (a) to "defense | pass" (b–c) to "defense | set" (d) and the inferred players' roles in each frame.

match the ground truth (Figure 11(c–d)). It is notable that such kind of error is inevitable, even for human experts who identify players' roles in a transitioning process without further information such as a player's name or jersey number, which is out of the scope of this article.

Action anticipation is performed using inferred team strategy and players roles, which is in accordance with Experiment 3 in Section 6.2. Anticipation results are shown in Figure 12–14 for two testing video clips with a framerate of 25 fps. Figure 12(a) shows that the setter, marked by the black bounding box, is predicted as the key player who will dominate the game based on the inferred role and team strategy. The observed action, the ground truth future action, and the anticipated action are visualized in the bar chart of Figure 12(b), and the red vertical line indicates where the current frame is temporally located in the testing sequence. More specifically, the first segment of the middle and bottom bar is of the same color as the top bar, representing that the current action would keep until the onset of the future action with a different color. The anticipation MLP gives the credible prediction of the key player who will be setting the ball, in spite of the discrepancy of 7 frames (0.28s) between the predicted timing and ground truth, as shown in the length of the middle and bottom bars (Figure 12(b)). Moreover, as time evolves from Figure 12(b) to 12(d), the difference in timing gradually reduces, indicating the update of anticipation result as the future unfolds.

On the other hand, more than one individuals can be predicted as key players, as shown in Figure 13, where the three key players are marked by the black bounding boxes. Based on a short observation sequence of 7 frames (0.28 s), the anticipation MLP predicts that both middle-hitter (player 8) and left-hitter (player 10) will launch a spiking, although the ground truth shows only the left-hitter eventually spikes the ball. Such mistake or conservatism is inevitable because it is yet uncertain in this moment who would launch the final attack as they both have great opportunity. This is also a general tactic when one of the hitters potentially makes a feint in order to distract

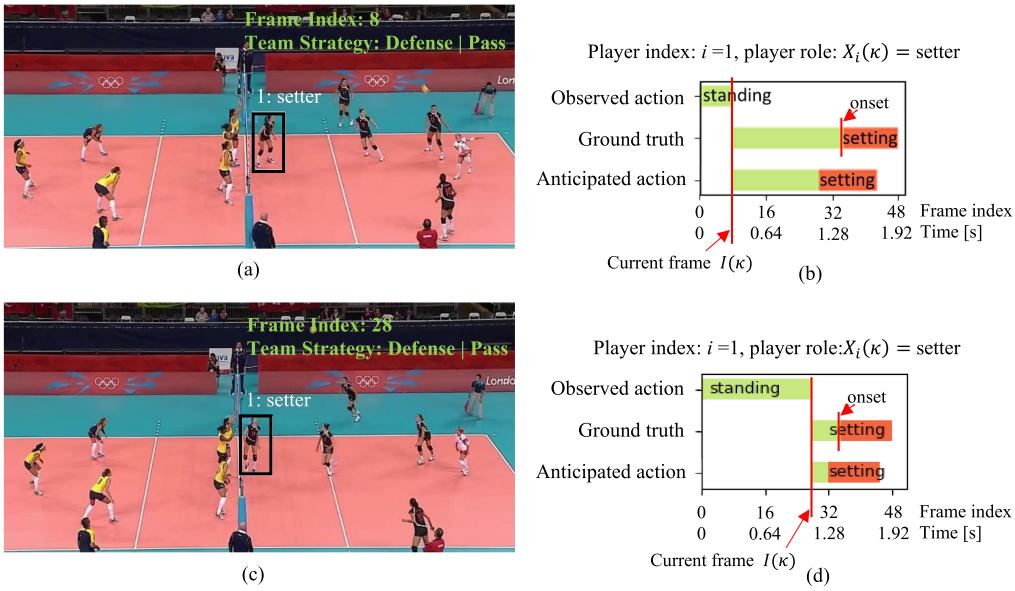


Fig. 12. The anticipated key player and action in the 8^{th} frame (a–b) and the 28^{th} frame (c–d) in a testing video clip.

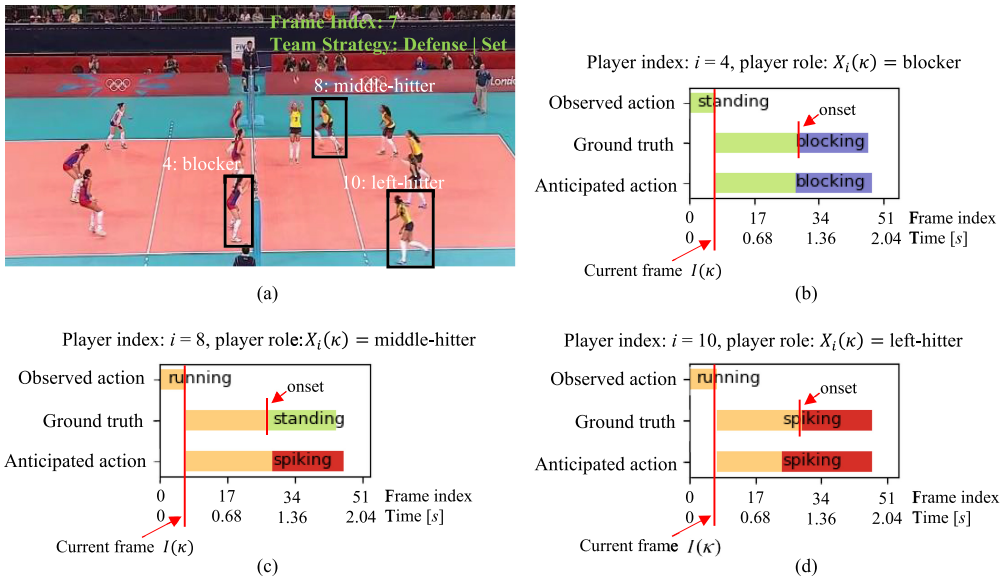


Fig. 13. Three key players (a) and the anticipated actions (b–d) in the 7^{th} frame of a testing video.

blockers of the opposing team. As the game proceeds, the anticipated action of the middle-hitter evolves, finally reaching to the ground truth, as illustrated in Figure 14(c). In addition, the onset and duration of the anticipated actions are indicated by the change of color and the length of the bars, respectively.

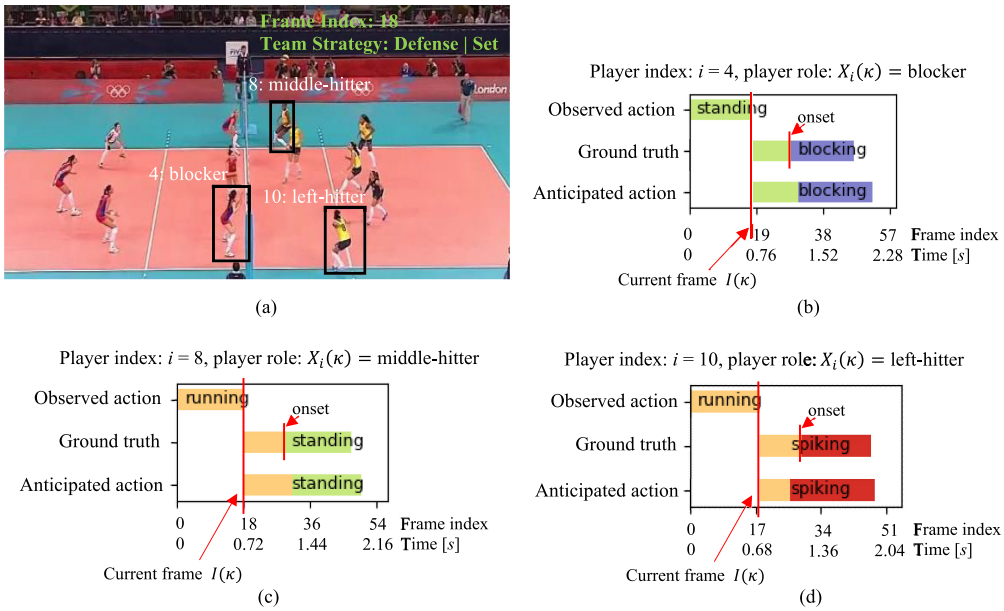


Fig. 14. Three key players (a) and the anticipated actions (b–d) in the 18^{th} frame of a testing video.

6.2 Performance Analysis and Results

The effectiveness of the inference and action anticipation algorithms presented in the previous sections is demonstrated using the metrics known as multi-class average precision (APr), multi-class average recall (ARc), and multi-class average accuracy (Ac). The APr score concerns the proportion of inferred values, consisting of both **true positive** (**TP**) and **false positive** (**FP**), that is actually true (i.e., APr = TP/(TP + FP)). In contrast, the ARc score is the proportion of ground truth labels, including both TP and **false negative** (**FN**), that is correctly inferred (i.e., ARc = TP/(TP + FN)). For both metrics, higher values correspond to better performance. Finally, Ac is defined as the harmonic mean of APr and ARc, which is also known as the F1-score

$$\text{Ac} = 2\,\frac{\text{APr} \times \text{ARc}}{\text{APr} + \text{ARc}}. \tag{32}$$

Two hidden variables, the team strategy ($S(\kappa)$) and the players' roles ($X(\kappa)$), are inferred in each frame with the overall results presented in Table 2. A comparative study is performed to assess the performance of the anticipation model as well as the robustness of the holistic framework, i.e., the dependence of the anticipating ability on the inferred hidden variables.

The comparative study involves three types of experiments aimed at determining the performance variability as a function of the hidden variables and corresponding inference accuracy:

— Experiment 1: perfect knowledge of team strategy ($S(\kappa)$) and player roles ($X(\kappa)$);
— Experiment 2: inferred team strategy ($S(\kappa)$) and perfect knowledge of player roles ($X(\kappa)$);
— Experiment 3: inferred team strategy ($S(\kappa)$) and player roles ($X(\kappa)$).

The purpose of the first experiment is to determine the performance of the action anticipation independently of the inference algorithm. The results in Table 2 show the important influence that the player role and team strategy have on the solution of the action anticipation problem (*problem 2*). As a result, the action anticipation performance degrades as errors are introduced

Table 2. Inference and Action Anticipation Performance

Experiment	Average Precision	Average Recall	Average Accuracy
Team strategy inference	0.87	0.82	84.43%
Role inference	0.88	0.86	86.99%
Experiment 1	0.92	0.89	90.47%
Experiment 2	0.88	0.86	86.99%
Experiment 3	0.81	0.80	80.50%

Table 3. Ablation Study Regarding the Hidden Role Variables

Model	Average Precision	Average Recall	Average Accuracy
Experiment 3	0.81	0.80	80.50%
Experiment 4	0.71	0.69	69.99%

in the inference stage, through Experiments 2 and 3. This is because, despite the excellent performance of the DMRF algorithm (Table 2), inferring the hidden variables from video introduces some errors (compared to perfect knowledge) that are, then, propagated to the action anticipation algorithm.

The advantage of this holistic approach is that action anticipation draws from the aggregation of both implicit hidden variables and explicit visual features. Therefore, errors from one source of information are potentially compensated by information obtained from other features. The performance results could be further improved by leveraging other variables and sensor modalities, which are easily incorporated in the proposed approach by augmenting the feature vectors. In addition, an ablation study is performed with a variant of the proposed model that excludes the inferred players' roles from the proposed holistic framework shown in Figure 4:

— Experiment 4: action anticipation without player roles ($X(\kappa)$) in the model input.

Results of Experiment 4 are compared against results of the holistic approach (Experiment 3) in Table 3. Without the knowledge of players' roles, Experiment 4 sees a significant drop in the action anticipation accuracy, which, by contrast, shows the improvement brought by the inference of hidden role variables to the solution of the action anticipation problem (*problem 2*).

The ability to predict the onset and duration of a future action is also critical, as well as coupled with the problem of anticipating the action type, since many algorithms assume the starting time is known or even observed. Team sports offer an excellent benchmark problem, because players constantly adjust the timing and duration of their actions, speeding up or slowing down actions and behaviors for strategic purposes. These difficulties are exacerbated by varying contexts, for example, because the trajectory of the ball and the skills of the opponents differ greatly from one team to another, yielding different samples in the training and testing datasets. The performance of action timing prediction is evaluated by the time-relative error, which is defined as the ratio of the absolute prediction error to the corresponding prediction horizon. Then, the **mean of the time-relative error** (**MTRE**) of each testing instance is used as the metric to assess the performance on the test database. The proposed model achieves an MTRE of 14.57% and 15.67% for the prediction of the action onset and duration, respectively. When compared to the LSTM solution proposed in [22] for anticipating an individual's cooking activity, the DMRF-MLP approach presented in this article achieves a comparable prediction horizon (0.48–1.84 s, versus 0.25–2 s) using a smaller observation time window (0.12–1.80 s, versus 1.75–3.50 s) and, thus, is applicable to fast actions and highly dynamic activities, such as sports.

7 CONCLUSION

This article presents a holistic approach that integrates image recognition, state estimation, and inference of hidden variables for the challenging problem of action anticipation in human teams. The approach is demonstrated on the team sport of volleyball, in which the team strategy and players' roles are unobservable and change significantly over time. The team strategy is first inferred by constructing a team feature descriptor that aggregates domain knowledge of volleyball games and features of individual players. Sequentially, the players' roles, modeled probabilistically as the DMRF graph, can be inferred using a MCMC sampling method. The dynamic graph structure that captures player interrelationships can be estimated by solving an integer linear program in each frame. By leveraging holistic information about the scene, including inferred team strategy, players' roles, as well as domain knowledge and instantaneous visual features, the action anticipation MLP is able to predict the semantic label and timing of the future actions by multiple interacting key players on the team. The numerical experiments show that this novel approach achieves an average accuracy of 84.43% for team strategy inference, 86.99% for role inference, and 80.50% for action anticipation. Additionally, the action onset and duration are predicted with a mean time-relative error of 14.57% and 15.67%, respectively.

REFERENCES

[1] Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I. Jordan. 2003. An introduction to MCMC for machine learning. *Machine Learning* 50, 1 (2003), 5–43.

[2] Anurag Arnab, Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Måns Larsson, Alexander Kirillov, Bogdan Savchynskyy, Carsten Rother, Fredrik Kahl, and Philip H. S. Torr. 2018. Conditional random fields meet deep neural networks for semantic segmentation: Combining probabilistic graphical models with deep learning for structured prediction. *IEEE Signal Processing Magazine* 35, 1 (2018), 37–52.

[3] Sina Mokhtarzadeh Azar, Mina Ghadimi Atigh, Ahmad Nickabadi, and Alexandre Alahi. 2019. Convolutional relational machine for group activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach, 7892–7901.

[4] Timur Bagautdinov, Alexandre Alahi, François Fleuret, Pascal Fua, and Silvio Savarese. 2017. Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, 4315–4324.

[5] Linchao Bao, Baoyuan Wu, and Wei Liu. 2018. CNN in MRF: Video object segmentation via inference in a CNN-based higher-order spatio-temporal MRF. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5977–5986.

[6] Murchana Baruah and Bonny Banerjee. 2020. A multimodal predictive agent model for human interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 1022–1023.

[7] Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning*. springer.

[8] Yang Cao, Yupin Luo, and Shiyuan Yang. 2011. Image denoising based on hierarchical Markov random field. *Pattern Recognition Letters* 32, 2 (2011), 368–374.

[9] Anirban Chakraborty and Amit K. Roy-Chowdhury. 2014. Context-aware activity forecasting. In *Proceedings of the Asian Conference on Computer Vision*. Springer, Singapore, 21–36.

[10] Michael Chan, Gabor T. Herman, and Emanuel Levitan. 1995. Bayesian image reconstruction using a high-order interacting MRF model. In *Proceedings of the International Conference on Image Analysis and Processing*. Springer, 608–614.

[11] Chao Chen, Shuhai Jiao, Shu Zhang, Weichen Liu, Liang Feng, and Yasha Wang. 2018. TripImputor: Real-time imputing taxi trip purpose leveraging multi-sourced urban data. *IEEE Transactions on Intelligent Transportation Systems* 19, 10 (2018), 3292–3304.

[12] Chao Chen, Qiang Liu, Xingchen Wang, Chengwu Liao, and Daqing Zhang. 2021. semi-Traj2Graph: Identifying fine-grained driving style with GPS trajectory data via multi-task learning. *IEEE Transactions on Big Data* (2021).

[13] Evan Cheshire, Cibele Halasz, and Jose Krause Perin. 2013. Player tracking and analysis of basketball plays. In *Proceedings of the European Conference of Computer Vision*.

[14] Siddhartha Chib and Srikanth Ramamurthy. 2010. Tailored randomized block MCMC methods with application to DSGE models. *Journal of Econometrics* 155, 1 (2010), 19–38.

[15] Kenneth James Williams Craik. 1952. *The Nature of Explanation*. Vol. 445. CUP Archive.

[16] Petar M. Djuric, Jayesh H. Kotecha, Jianqui Zhang, Yufei Huang, Tadesse Ghirmai, Mónica F. Bugallo, and Joaquin Miguez. 2003. Particle filtering. *IEEE Signal Processing Magazine* 20, 5 (2003), 19–38.
[17] Junyi Dong, Pingping Zhu, and Silvia Ferrari. 2020. Oriented pedestrian social interaction modeling and inference. In *Proceedings of the 2020 American Control Conference*. IEEE, Virtual, 1373–1370.
[18] Nour Eldin Elmadany, Yifeng He, and Ling Guan. 2021. Improving action recognition via temporal and complementary learning. *ACM Transactions on Intelligent Systems and Technology* 12, 3 (2021), 1–24.
[19] Dirk Farin, Susanne Krabbe, Wolfgang Effelsberg, et al. 2003. Robust camera calibration for sport videos using court models. In *Proceedings of the Storage and Retrieval Methods and Applications for Multimedia 2004*, Vol. 5307. International Society for Optics and Photonics, 80–91.
[20] Alircza Fathi, Jessica K. Hodgins, and James M. Rehg. 2012. Social interactions: A first-person perspective. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Providence, 1226–1233.
[21] Silvia Ferrari and Thomas A. Wettergren. 2021. *Information-Driven Planning and Control*. MIT Press.
[22] Antonino Furnari and Giovanni Maria Farinella. 2019. What would you expect? Anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Long Beach, 6252–6261.
[23] Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. 2019. Predicting the future: A jointly learnt model for action anticipation. In *Proceedings of the IEEE International Conference on Computer Vision*. Long Beach, 5562–5571.
[24] Stuart Geman and Donald Geman. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6, 6 (1984), 721–741.
[25] Anthony Giddens and Philip W. Sutton. 2021. *Essential Concepts in Sociology*. John Wiley & Sons.
[26] Josep M. Gonfaus, Xavier Boix, Joost Van de Weijer, Andrew D. Bagdanov, Joan Serrat, and Jordi Gonzalez. 2010. Harmony potentials for joint classification and segmentation. In *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, San Francisco, 3280–3287.
[27] Ankur Gupta, James J. Little, and Robert J. Woodham. 2011. Using line and ellipse features for rectification of broadcast hockey video. In *Proceedings of the 2011 Canadian Conference on Computer and Robot Vision*. IEEE, St Johns, 32–39.
[28] Fasih Haider, Fahim A. Salim, Dees B. W. Postma, Robby Van Delden, Dennis Reidsma, Bert-Jan van Beijnum, and Saturnino Luz. 2020. A super-bagging method for volleyball action recognition using wearable sensors. *Multimodal Technologies and Interaction* 4, 2 (2020), 33.
[29] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*. Venice, 2961–2969.
[30] De-An Huang and Kris M. Kitani. 2014. Action-reaction: Forecasting the dynamics of human interaction. In *Proceedings of the European Conference on Computer Vision*. Springer, Zurich, 489–504.
[31] Mostafa S. Ibrahim and Greg Mori. 2018. Hierarchical relational networks for group activity recognition and retrieval. In *Proceedings of the European Conference on Computer Vision*. Munich, 721–736.
[32] Mostafa S. Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. 2016. A hierarchical deep temporal model for group activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, 1971–1980.
[33] Luca Iocchi. 2006. Robust color segmentation through adaptive color distribution transformation. In *Proceedings of the Robot Soccer World Cup*. Springer, 287–295.
[34] Thorsten Joachims. 2006. Structured output prediction with support vector machines. In *Proceedings of the Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition and Structural and Syntactic Pattern Recognition*. Springer, Hong Kong, 1–7.
[35] Qiuhong Ke, Mario Fritz, and Bernt Schiele. 2019. Time-conditioned action anticipation in one shot. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Long Beach, 9925–9934.
[36] Davis E. King. 2009. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research* 10, 7 (2009), 1755–1758.
[37] Daphne Koller and Nir Friedman. 2009. *Probabilistic Graphical Models: Principles and Techniques*. MIT press.
[38] Tian Lan, Tsung-Chuan Chen, and Silvio Savarese. 2014. A hierarchical representation for future action prediction. In *Proceedings of the European Conference on Computer Vision*. Springer, Zurich, 689–704.
[39] Kang Li and Yun Fu. 2014. Prediction of human activity by discovering temporal sequence patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 8 (2014), 1644–1657.
[40] Ralph Linton. 1936. The study of man: An introduction. D. Appleton-Century company, incorporated.
[41] Tahmida Mahmud, Mahmudul Hasan, and Amit K. Roy-Chowdhury. 2017. Joint prediction of activity labels and starting times in untrimmed videos. In *Proceedings of the IEEE International Conference on Computer Vision*. Venice, 5773–5782.

[42] Partha Pratim Mondal, Giuseppe Vicidomini, and Alberto Diaspro. 2007. Markov random field aided Bayesian approach for image reconstruction in confocal microscopy. *Journal of Applied Physics* 102, 4 (2007), 044701.
[43] M. Naveenkumar and S. Domnic. 2020. Deep ensemble network using distance maps and body part features for skeleton based action recognition. *Pattern Recognition* 100 (2020), 107125.
[44] Lukas Neumann, Andrew Zisserman, and Andrea Vedaldi. 2019. Future event prediction: If and when. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. Long Beach, 0–0.
[45] Sebastian Nowozin, Christoph H. Lampert 2011. Structured learning and prediction in computer vision. *Foundations and Trends® in Computer Graphics and Vision* 6, 3–4 (2011), 185–365.
[46] Vignesh Ramanathan, Bangpeng Yao, and Li Fei-Fei. 2013. Social role discovery in human events. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Portland, 2475–2482.
[47] João Ramos, Rui J. Lopes, and Duarte Araújo. 2018. What's next in complex networks? Capturing the concept of attacking play in invasive team sports. *Sports Medicine* 48, 1 (2018), 17–28.
[48] João Ribeiro, Keith Davids, Duarte Araújo, Pedro Silva, João Ramos, Rui Lopes, and Júlio Garganta. 2019. The role of hypernetworks as a multilevel methodology for modelling and understanding dynamics of team sports performance. *Sports Medicine* 49, 9 (2019), 1337–1344.
[49] Cristian Rodriguez, Basura Fernando, and Hongdong Li. 2018. Action anticipation by predicting future dynamic images. In *Proceedings of the European Conference on Computer Vision*. Munich, 1–10.
[50] Mohammad Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Basura Fernando, Lars Petersson, and Lars Andersson. 2017. Encouraging lstms to anticipate actions very early. In *Proceedings of the IEEE International Conference on Computer Vision*. Venice, 280–289.
[51] Paul Schnitzspan, Mario Fritz, Stefan Roth, and Bernt Schiele. 2009. Discriminative structure learning of hierarchical representations for object detection. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Miami, 2238–2245.
[52] Yuge Shi, Basura Fernando, and Richard Hartley. 2018. Action anticipation with rbf kernelized feature mapping rnn. In *Proceedings of the European Conference on Computer Vision*. Munich, 301–317.
[53] Tianmin Shu, Sinisa Todorovic, and Song-Chun Zhu. 2017. Cern: Confidence-energy recurrent network for group activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, 5523–5531.
[54] Chenyang Si, Ya Jing, Wei Wang, Liang Wang, and Tieniu Tan. 2020. Skeleton-based action recognition with hierarchical spatial reasoning and temporal stack learning network. *Pattern Recognition* 107 (2020), 107511.
[55] Khurram Soomro, Haroon Idrees, and Mubarak Shah. 2018. Online localization and prediction of actions and interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2018), 459–472.
[56] Yansong Tang, Jiwen Lu, Zian Wang, Ming Yang, and Jie Zhou. 2019. Learning semantics-preserving attention and contextual interaction for group activity recognition. *IEEE Transactions on Image Processing* 28, 10 (2019), 4997–5012.
[57] Rajkumar Theagarajan, Federico Pala, Xiu Zhang, and Bir Bhanu. 2018. Soccer: Who has the ball? Generating visual analytics and player statistics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. Salt Lake City, 1749–1757.
[58] Henry L. Tischler. 2013. *Cengage Advantage Books: Introduction to Sociology*. Cengage Learning.
[59] Xiaofeng Tong, Jia Liu, Tao Wang, and Yimin Zhang. 2011. Automatic player labeling, tracking and field registration and trajectory mapping in broadcast soccer video. *ACM Transactions on Intelligent Systems and Technology* 2, 2 (2011), 1–32.
[60] A. Vedaldi and B. Fulkerson. 2008. VLFeat: An open and portable library of computer vision algorithms. Retrieved from http://www.vlfeat.org/.
[61] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. 2016. Anticipating visual representations from unlabeled video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, 98–106.
[62] Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert. 2016. An uncertain future: Forecasting from static images using variational autoencoders. In *Proceedings of the European Conference on Computer Vision*. Springer, 835–851.
[63] Suhang Wang, Charu Aggarwal, and Huan Liu. 2018. Random-forest-inspired neural networks. *ACM Transactions on Intelligent Systems and Technology* 9, 6 (2018), 1–25.
[64] Yingying Wang, Qingchun Ji, and Chenglin Zhou. 2019. Effect of prior cues on action anticipation in soccer goalkeepers. *Psychology of Sport and Exercise* 43 (2019), 137–143.
[65] Qiong Wu and Pierre Boulanger. 2016. Enhanced reweighted MRFs for efficient fashion image parsing. *ACM Transactions on Multimedia Computing, Communications, and Applications* 12, 3 (2016), 1–16.
[66] Zhirong Wu, Dahua Lin, and Xiaoou Tang. 2016. Deep markov random field for image modeling. In *Proceedings of the European Conference on Computer Vision*. Springer, Amsterdam, 295–312.

[67] Rui Yan, Jinhui Tang, Xiangbo Shu, Zechao Li, and Qi Tian. 2018. Participation-contributed temporal dynamic model for group activity recognition. In *Proceedings of the 26th ACM International Conference on Multimedia*. Seoul, 1292–1300.

[68] Shengping Zhang, Hongxun Yao, Xin Sun, and Shaohui Liu. 2012. Robust visual tracking using an effective appearance model based on sparse coding. *ACM Transactions on Intelligent Systems and Technology* 3, 3 (2012), 1–18.

[69] Yu Zhu, Wenbin Chen, and Guodong Guo. 2015. Fusing multiple features for depth-based action recognition. *ACM Transactions on Intelligent Systems and Technology* 6, 2 (2015), 1–20.

Received December 2021; revised March 2022; accepted April 2022