IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 15 September 2023, accepted 15 November 2023, date of publication 23 November 2023,
date of current version 29 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3336425

RESEARCH ARTICLE

Action Valuation of On- and Off-Ball Soccer Players Based on Multi-Agent Deep Reinforcement Learning

HIROSHI NAKAHARA[1], KAZUSHI TSUTSUI[1], KAZUYA TAKEDA[1], (Senior Member, IEEE), AND KEISUKE FUJII[1,2,3], (Member, IEEE)
[1]Graduate School of Informatics, Nagoya University, Furo-cho, Chikusa, Nagoya, Aichi 464-8601, Japan
[2]RIKEN Center for Advanced Intelligence Project, Yamadaoka, Suita, Osaka 103-0027, Japan
[3]JST PRESTO, Chiyoda, Tokyo 102-0076, Japan

Corresponding author: Keisuke Fujii (fujii@i.nagoya-u.ac.jp)

This work was supported in part by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant 20H04075 and Grant 21H05300, and in part by the Japan Science and Technology Agency (JST) Precursory Research for Embryonic Science and Technology (PRESTO) under Grant JPMJPR20CA.

ABSTRACT Analysis of invasive sports such as soccer is challenging because the game situation changes continuously in time and space, and multiple agents individually recognize the game situation and make decisions. Previous studies using deep reinforcement learning have often considered teams as a single agent and valued the teams and players who hold the ball in each discrete event. Then it was challenging to value the actions of multiple players, including players far from the ball, in a spatiotemporally continuous state space. In this paper, we propose a method of valuing possible actions for on- and off-ball soccer players in a single holistic framework based on multi-agent deep reinforcement learning. We consider a discrete action space in each data frame that mimics that of Google research football and leverages supervised learning for actions in reinforcement learning. In the experiment, we analyzed the relationships with conventional indicators, season goals, and game ratings by experts, and showed the effectiveness of the proposed method. Our approach can assess how multiple players move continuously throughout the game, which is difficult to be discretized or labeled but vital for teamwork, scouting, and fan engagement.

INDEX TERMS Multi-agent, reinforcement learning, sports, football.

I. INTRODUCTION

With advanced data analysis in sports, tactical planning, player evaluation, and coaching methods based on data have recently become available. The data analytics of dynamic movements in team sports such as soccer is considered challenging because game situations are continuous in time and space, and multiple agents individually recognize the game situations and make decisions. In particular, each action's value has been quantified based on the strength of its association with scores for on-ball players (i.e., with a ball). In previous work using supervised learning, machine learning models were used to compute an action's value by predicting whether scores or other events occur or not in the following actions [1], [2], [3], [4]. In these frameworks, it would be difficult to consider possible (i.e., counterfactual) actions as time goes back from a goal or other events. To value on-ball actions in terms of obtaining rewards (e.g., goals), there have been studies using reinforcement learning (RL) [5], [6], [7], [8]. These works typically consider teams as a single agent and valuate an on-ball player or a team in irregularly occurring events (e.g., passes and shots). Considering the nature of soccer, it should be modeled in each data frame, and valuate actions without event labels including on- and off-ball plays.

Although most studies focus on the evaluation of on-ball players, players can indirectly contribute to scoring even when they are off-ball (i.e., without a ball) which is a large part of their playing time (e.g., approximately 87 min of 90 min [9]). A previous work [10] estimated

The associate editor coordinating the review of this manuscript and approving it for publication was Muhammad Asif.

the value of the state from the positional information of the ball and players based on a rule-based model called Off-Ball Scoring Opportunities (OBSO). However, valuing other attacking players who do not receive the ball and reflecting the valuation of several possible actions in the state value is challenging. Another study [11] quantified every off-ball player's impact on scores in terms of the difference between predicted and real player movements. The method quantitatively values only a single player's contribution once through their predicted movement trajectories. Thus it is challenging to calculate the contributions of multiple players at each time stamp using a comprehensive learning-based framework.

In this paper, we propose a valuation method of on- and off-ball soccer players in a single holistic framework based on multi-agent deep RL. Specifically, we consider a discrete action space in each data frame that mimics that of the Google research football (GFootball) [12] and leverage supervised learning for actions in RL. Based on a deep RL model with a discrete action space, we estimate the possible action value of multiple players in real games, including those far from the ball. The proposed network estimates state-action values (i.e., Q-values) based on the game states (e.g., player and ball locations) and actions (e.g., shot and pass). For the off-ball action, we define the directions of movement as actions. For player valuations, our method can compute the overall contribution of each player during the attack by aggregating the Q-values in the RL model. The proposed method enables us to comprehensively value on- and off-ball actions, which makes it possible to compare the contributions of the two types of actions, and thus provides important information for understanding the characteristics of players. Assessing the movements of all players for teammates is important for building teamwork, assessment of players' salaries, recruitment, and scouting.

In summary, our main contributions were as follows. (1) We propose a valuation method of on- and off-ball soccer players in a single holistic framework based on multi-agent deep RL, which contributes to the interdisciplinary field between sports analytics and machine learning. (2) We consider a discrete action space in each data frame that mimics GFootball and leverages supervised learning for actions in RL to restrict the state and action spaces from the limited data. Furthermore, since the proposed method can counterfactually compute possible action values that were not chosen in the actual game, it can be also used for valuating counterfactual actions. (3) In the experiment, we analyzed the relationships with conventional indicators, season goals, and game ratings by experts, and showed the effectiveness of the proposed method. Our approach can assess how multiple players move continuously throughout the game, which is difficult to be discretized or labeled but vital for teamwork, scouting, and fan engagement. The remainder of this paper is structured as follows. First, we overview the related works in Section II and describe our methods in Section III. Next, we present experimental results in Section IV and conclude this paper in Section V.

II. RELATED WORK

In team sports tactical behaviors, agents' behavioral process bears resemblance to the RL framework (e.g., [13]). Due to various challenges in modeling the entire framework from data, two approaches can be adopted: estimating related variables and functions from data as a sub-problem (i.e., the inverse approach) and constructing a model (e.g., an RL model) to generate data in a virtual environment (i.e., the forward approach such as [12], [14]). In this paper, we have concentrated on the inverse approach but also considered a forward model.

Numerous approaches have been employed to quantitatively assess the actions of attacking players in terms of scoring, such as using expected scores derived from tracking data [7], [15], [16], [17], [18], and action data such as dribbling and passing [1], [4], [19]. Some researchers have valuated passes [20], [21], [22], while others have assessed actions to receive a ball by attributing a value to the location with the highest expected score [10], [23] and using a rule-based approach [24]. Defensive behaviors have also been valued using data-driven methods [2], [25]. However, such approaches would have difficulties in considering possible (i.e., counterfactual) actions as time goes back from a goal or other events.

From the RL perspective, numerous studies have focused on inverse approaches. To value on-ball actions, several studies have estimated Q-function or other policy functions [5], [6], [8], [17]. However, they often consider teams as a single agent and did not valuate off-ball players in all time steps (without events). In terms of inverse RL, research on estimating reward functions has also been conducted [26], [27]. To estimate policy functions, researchers have sometimes performed trajectory prediction through imitation learning [28], [29], [30], [31] and behavioral modeling [32], [33], [34], [35], aiming to mimic (rather than optimize) a policy using neural networks. Our approach considers the RL model overall rather than as a sub-problem using hand-crafted reward functions, and estimates multiple players' Q-functions for simultaneously valuating on- and off-ball players even when no event occurs.

III. METHOD

In this section, we describe our problem setting and propose our RL framework. Then we describe the dataset and preprocessing, and valuation framework.

A. PROBLEM SETTING

In this study, we aim to value players by computing state-action values (i.e., Q-values) using an RL framework. For simplicity, here we consider independent multi-agent RL (but rewards are shared like actual soccer; then implicitly cooperative) and then omit the agent index.

The RL model in this study consists of three components: state s, action a, and reward r. At the time t, when an action a_t is chosen in a state s_t, a reward r_{t+1} is given. In on-policy RL, the agent learns a policy π that can maximize $\sum_{t=1}^{T} \gamma^t r_t$, where $\gamma \in [0, 1]$ is the discount factor and T is the time horizon. We consider an episode in RL as possessions described later, and usually all possesesions are not a fixed length. Hereafter, we consider the case of $\gamma = 1$ [6] with no discount for simplicity.

For the state s_t, we used 2-dimensional position coordinates and velocities data of the players (22) and the ball (23 × 2 × 2 dimensions). Inspired by 19 actions defined in GFootball [12], we selected 14 actions of the attacking players: movement actions defined as 8 different movement directions (8 directions in 45 degree increments), idle, starting and stopping a sprint, the release of the movement direction, passing, and shooting. The passing and shooting actions were defined by the labels (event data) in the dataset described later. Movement directions were computed based on the player's velocity direction. Other, idle, starting and stopping a sprint, and release of the movement direction were computed based on the stopping (0.1 m/s) and sprint thresholds (24 km/h). Regarding on-ball players' actions, the dataset included other behavior labels (e.g., dribbling and trapping), which were not used in this study for simplicity. Note that we consider the agents can fully observe their state and that an off-ball player's on-ball action is ignored in the post-hoc processing for simplicity based on the original GFootball setting [12]. These points should be improved in future work.

For the reward r, we added the following three values at the last time T of the sequence of attacks. Usually the reward should be only goals in team sports but it is generally difficult to learn the model based on sparse rewards. Then we consider the following three rewards: (1) Goal: 1 if a series of attacks end in a score, and 0 otherwise, (2) Expected score with no goal: the EPV (expected possession value) [36], which is an indicator of the probability of scoring based on the ball position coordinates (x, y) at time T, (3) Conceding: −1 if a goal is scored by an opponent's attack immediately after a sequence of attacks, and 0 otherwise. For the rewards at t other than time T $(t < T)$, we used 0 in this study. We used the EPV based on the data at https://github.com/Friends-of-Tracking-Data-FoTD/LaurieOnTracking.

B. DEEP REINFORCEMENT LEARNING WITH ACTION SUPERVISION

Next, we explain our RL algorithm with action supervision. We compute state-action values (Q-value) for the valuation of plays, which has been used in previous sports studies [5], [6], [8], [37]. In this framework, the optimal state-action value function $Q^*(s_t, a_t)$ is determined by solving the Bellman equation as follows:

$$Q^*(s_t, a_t) = E_{s_{t+1},a_{t+1}}[r_{t+1} + Q(s_{t+1}, a_{t+1})|s_t, a_t]$$

$$= \sum_{r_{t+1}} P(r_{t+1}|s_t, a_t)r_{t+1} + \sum_{s_{t+1}a_{t+1}} P(s_{t+1}, a_{t+1}|s_t, a_t)Q(s_{t+1}, a_{t+1}), \tag{1}$$

where $Q(s_t, a_t)$ refers to the Q-value of taking action a in state s. In other words, the Q-value of taking action a_t in state s_t is defined as the sum of the expected reward when the sequence of attacks ends at time t and the expected Q-value at time $t + 1$.

To train the RL model, following by [5] and [6], we adopt a temporal difference (TD) prediction method called SARSA (state-action-reward-state-action), which is a model-free and on-policy RL algorithm, with recurrent neural network (RNN). As the input sequences, we used a matrix of T frames of state (92 dim.) and the one-hot action for each agent (14 dim.) as input. A conceptual diagram of the architecture is shown in Fig. 1. We used single-layer GRUs (gated recurrent units) as RNNs with 64 hidden neurons. The input layer was a single layer with 64-dimensional neurons and the activation function was ReLU. In the output layer, we output the Q-values for each player for each action (14 dim.). We consider a simple RNN model due to the small sample size and the number of total learnable parameters is 32,718. To estimate Q values based on Eq. (1), we compute the following TD loss:

$$\mathcal{L}_{TD} = \sum_{t \in T} (r_{t+1} + Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t))^2. \tag{2}$$

In addition, we introduce a supervised loss for actions because the above approach might lead to insufficient learning of all possible actions and it would be necessary to restrict the state spaces from the limited data. Based on the discussion in [38] and [39], we propose a simple action supervision loss represented by the cross-entropy of softmax values of the Q-function such that

$$\mathcal{L}_{AS} = -\sum_{t \in T} \mathbf{a}_t \cdot \log\left(\text{softmax}\left(\mathbf{Q}_{s_t}\right)\right), \tag{3}$$

where $\mathbf{a}_t \in \{0, 1\}^{|A|}$ (i.e., one-hot vector of actions), $|A|$ is the size of action space, $\mathbf{Q}_{s_t} = [Q(s_t, a_t = 1), \ldots, Q(s_t, a = |A|)]$, and the log applies element-wise. This loss aims to maximize the Q-function values for the action of the data. Note that we assume that the action of the actual players is better than random selection. This is an inductive bias for efficient learning to estimate Q-values for all possible actions but the main loss is $\mathcal{L}_{TD}$ and the weight of $\mathcal{L}_{AS}$ should be much smaller than $\mathcal{L}_{TD}$.

We also add an L_1 regularization loss applied to the weights and biases of the network to help prevent over-fitting on the relatively small demonstration dataset. The total loss in our training is a combination of three losses:

$$\mathcal{L}_{total} = \mathcal{L}_{TD} + \lambda_1 \mathcal{L}_{AS} + \lambda_2 \mathcal{L}_{L_1}. \tag{4}$$

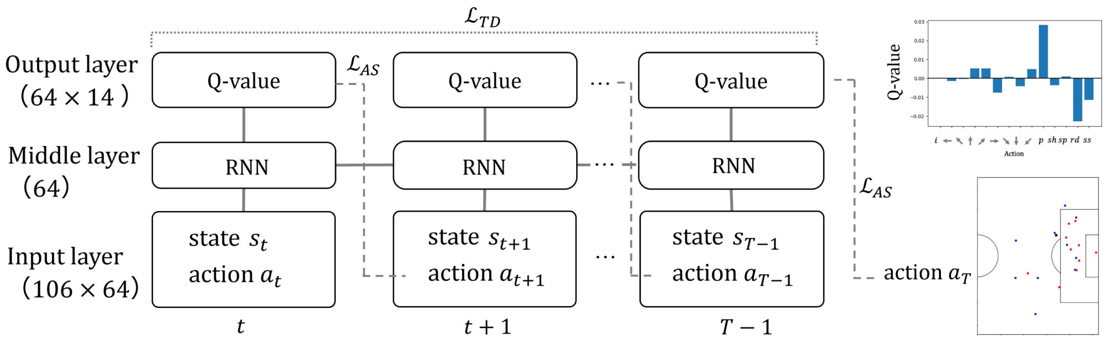


FIGURE 1. Our deep RL model with action supervision. We consider independent multi-agent RL (but rewards are shared like actual soccer) and then the model is constructed per player. The inputs of the network are the 92 dimensional state including position and velocity for 22 players and the ball, and 14 dimensional one-hot action. In the input layer, the input information is transformed into 64-dimensional hidden states for RNN (middle layer). Finally, Q-values for each action (i.e., 14 dimensions) are computed via the output layer.

We set $\lambda_1 = 0.001$ and $\lambda_2 = 0.0001$, which control the weighting among the losses. We determined the weight based on the scales of each loss in the validation phase such that the TD loss is not much smaller than other losses. We consider that the generalization performance based on the loss function is not directly linked to the goodness as the valuation indicators of soccer players. In addition to the limited data, for this reason, we did not perform intensive experiments using various deep learning models and hyperparameters. We trained the models using the Adam optimizer [40] with default parameters.

C. DATASET AND PREPROCESSING

In this study, we used 55 games data in the Meiji J1 League (a professional soccer league in Japan) 2019 season including all 34 games data of Yokohama F Marinos to perform specific player-level evaluations in limited data. Note that currently the tracking data for all players and timesteps were not publicly shared in such amounts and compared to the previous work [5], [6], [8] with only event data, we used tracking data with every 0.1 s, and then we had more samples for each game as described below. The dataset includes event data (i.e., labels of actions, e.g., passing and shooting, recorded at 30 Hz and the simultaneous xy coordinates of the ball) and tracking data (i.e., xy coordinates of all players recorded at 25 Hz) provided by Data Stadium Inc. The season goals for each player in each match were collected from [41]. The ratings by experts in each match [42] were also used, which were scored in 0.5-point increments with a maximum of 10 points.

As a preprocessing step, we first split a sequence of a game (usually about 90 minutes) into possessions (i.e., sequences of attacks) from the beginning of the possession (ball recovery) to the end of the possession (ball loss or goal) as an input of RL models. The minimum number of frames was 50, and the maximum number of frames was $T = 300$ (equivalent to 30 seconds) based on the insight of [43]. We performed zero-padding if the length of the possessions was less than T and did not perform back-propagation about the zero-padding frames. The tracking data were down-sampled to 10 Hz based on [11] and [31]. We analyzed 10 attacking players (without a goalkeeper), i.e., constructed 10 RL agent models. For the player assignment problem, we simply assigned the 10 players in the mean locations in a possession. That is, we assigned player 0 in the rightmost attacker on average in the possession and assigned player 10 in the leftmost attacker (except for the goalkeeper).

D. VALIDATION OF THE MODEL AND VALUATION OF PLAYERS

In the experiment, we first validated the RL model and then analyzed the valuations of professional soccer players. For simplicity, only attacks within the attacking third were used. To train the model, we used attack sequences other than those of Yokohama F. Marinos (1652 sequences) as the training data (includes 10% validation data). To value the players, we used attack sequences of Yokohama F. Marinos (1176 series) as the test data. The number of frames used in the training and test were 257,164 and 212,838 frames, respectively. Totally, we analyzed 14 players in Yokohama F. Marinos with enough playing time and note that we cannot use the trained model for each player because we used only the opponent data (i.e., we used the same assignment rules for Yokohama's team and 10 RL agent models were not assigned as each player in Yokohama. We efficiently utilized the limited data by modeling ten players based on the mean location in each possession and small and simple RNN-based RL model with action supervision. Based on these facts, we believe that the size of the dataset would be enough for our task.

IV. RESULTS

In this section, we validated the RL model, and then analyzed the valuations of professional soccer players. We quantitatively compared the test losses of various hyperparameters in the former, but in the latter, since there is no ground truth in the evaluation metric, then we qualitatively

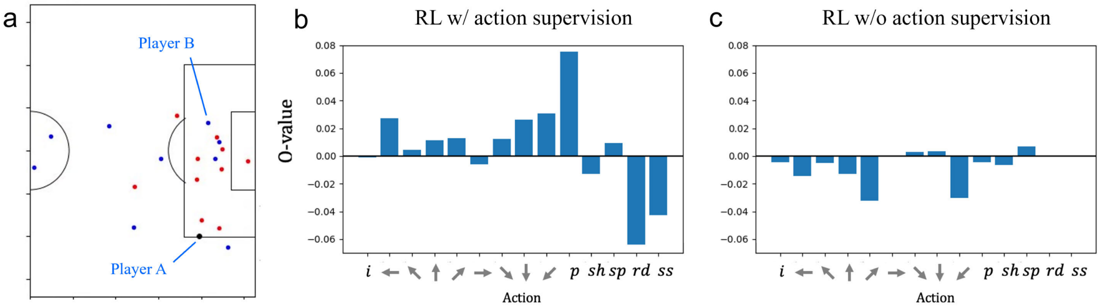


FIGURE 2. Example of estimated Q-values. (a) shows an example attack. Blue, red, and black indicate the attacker, defender, and the ball, respectively (note that the ball is over a defender). (b) and (c) show the Q-value of player A for each action using the RL model with and without action supervision, respectively. Note that *i*, *p*, *sh*, *sp*, *rd*, *ss* correspond to idle, pass, shot, sprint, decelerate, and sprint end, respectively, and arrows correspond to the direction of movement.

TABLE 1. Performances of our approach in terms of three loss functions and Q-values in on- and off-ball situations.

	w/ action supervision		w/o action supervision	
	w/ L_1	w/o L_1	w/ L_1	w/o L_1
$\mathcal{L}_{TD}$	0.00056 ± 0.00003	0.00053 ± 0.00004	0.00053 ± 0.00003	0.00045 ± 0.00003
$\mathcal{L}_{AS}$	3.8839 ± 0.00045	3.7371 ± 0.0015	3.9588 ± 0.00020	3.9585 ± 0.00074
$\mathcal{L}_{L_1}$	386.62	1737.59	386.80	1734.82
$Q_{\text{on-ball}}$	0.07052 ± 0.00090	0.06054 ± 0.00107	-0.00240 ± 0.00010	-0.00202 ± 0.00037
$Q_{\text{off-ball}}$	0.00667 ± 0.00757	0.00279 ± 0.00949	0.00100 ± 0.00720	0.00061 ± 0.00685

evaluate and quantitatively compared our approach with the existing metrics. Note that our aim is not to show the best model with the lowest test losses but to show the effectiveness of our new approach in a completely new problem setting to evaluate both on- and off-ball soccer players in the same framework.

A. VALIDATION OF OUR RL MODEL

Here we compared test losses in our RL models with and without action supervision (i.e., $\lambda_1 = 0$) and with and without L_1 loss (i.e., $\lambda_2 = 0$) in Eq. (4). We regard the models with action supervision as our proposed method and those without the supervision as the baselines, respectively. Due to the limited data (the reason is described in Section III-C), we did not perform intensive experiments using more sophisticated deep learning models. We quantitatively compared the mean and standard deviation of the TD loss $\mathcal{L}_{TD}$ in RL, action supervision loss $\mathcal{L}_{AS}$, and L_1 regularization loss in test samples as shown in Table 1. Regarding the losses in Table 1 (upper), the $\mathcal{L}_{TD}$s with action supervision (our methods) were larger than those without supervision (baselines), but $\mathcal{L}_{AS}$s with the supervision were smaller than that without supervision. Considering the fact that the reward scale is $[-1, 1]$, the learning of Q-values was considered to be within the acceptable range for both models even in the training of 55 games (again, compared with the previous work [5], [6], [8] with only event data, we used tracking data with every 0.1 s, and then we had more samples for each game). It should be noted that we can quantitatively compare the optimization results (i.e., losses) but can only qualitatively investigate the effectiveness of the model in terms of modeling soccer agents. In particular, the action supervision may require careful interpretation, because much less supervision (e.g., $\lambda_1 \ll 0.001$) would lead to insufficient learning of counterfactual action values, whereas much more supervision (e.g., $\lambda_1 \gg 0.001$) may overfit to the actual actions and would not consider counterfactual actions. Then we carried out a qualitative analysis as described below.

Next, an example of the Q-value output is shown. Fig. 2a shows the coordinates of the player and the ball in the frame to be valuated. In this case, player A in the actual game passed the ball to player B, who was considered to have more space for a shot. Figs. 2b and c show the Q-value of player A for each action using the RL model with and without action supervision, respectively. As shown in Fig. 2b, our model with action supervision indicates that the Q-value of a pass was higher than those of others, suggesting that passing may produce a more favorable result rather than other actions (e.g., a shot) in this case. In contrast, in Fig. 2c, our model without action supervision shows more distributed Q-values closer to zero and small value of the pass action. Quantitatively, we averaged Q-values of the observed actions among all analyzed frames and possessions. In Table 1 (lower), the average on-ball and off-ball Q-values with supervision were larger than those without supervision, and in action supervision models, those with L_1 regularization were also larger than those without L_1 regularization. In particular, our approach with supervision emphasized the on-ball valuation, which was similar to the conventional valuation of the players, but in some cases, a fairer valuation including off-ball situations may be required. Although which models were best in a practical sense cannot be determined from the data, for validation of the model output, we mainly show the results of the model with action supervision and L_1 regularization based on the above verification.

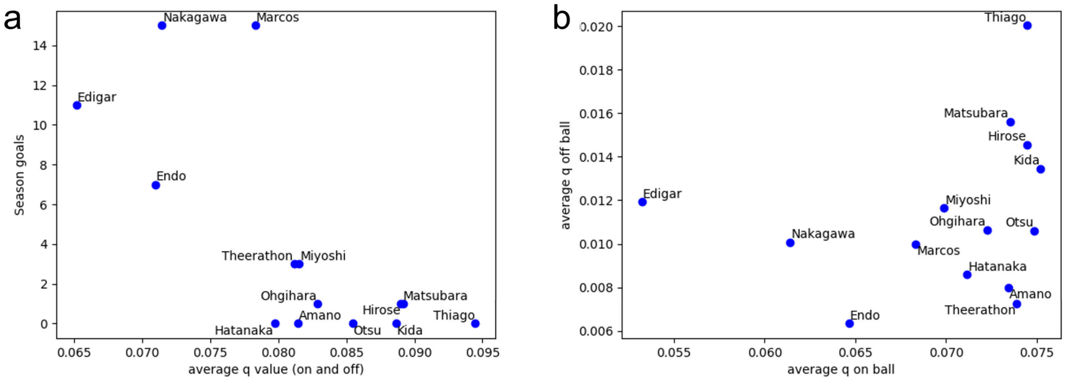


FIGURE 3. **The relationship between (a) average Q-values computed by the proposed method and season goals and (b) average Q-values of on- and off-ball states.**

B. COMPARISON WITH CONVENTIONAL INDICATORS AND PLAYER VALUATION

Since there is no true value for the Q-value, the performance of the model cannot be directly validated. Here we show the usefulness of the proposed indicators by investigating the relationship with the existing indicators. Specifically, we explain the relationships between the average Q-values computed by the proposed method and four indicators: season goals, ratings by experts, OBSO [10], and creating off-ball scoring opportunity (C-OBSO) [11]. OBSO was used to estimate the value of the state based on the rule-based model, which basically values attacking players who will receive the ball. C-OBSO quantifies the off-ball player's impact on scores created by the difference between predicted and real player movements. Spearman's rank correlation coefficient (hereafter denoted as ρ) was used as the correlation coefficient, and the players to be valuated were limited to those who played at least 10 games. Since the sample size was small ($N = 14$) in the correlation analysis, the ρ value was used as an effect size for evaluation (criteria are based on [44]), rather than the p-value.

First, the relationship between the season goals and the average Q-values obtained by the proposed method is shown in Fig. 3a. There was a moderate negative correlation between them ($\rho = -0.665$) whereas many players had 0 or 1 total goal. In the season, four players (Nakagawa, Marcos, Thiago, and Kida) were selected as the best 11 players award in the league. Among these four players, the proposed method showed higher values for a defender (Thiago), midfielder (Kida), and forward players (Marcos and Nakagawa) in this order. Considering the high negative correlation, our indicator tended to value defensive players who provided many passes, because Thiago ranked 6th in terms of the players with the most passes in the league in spite of being a defender. Similarly, Hatanaka, who was a defender ranked 1st in our indicator, also ranked 2nd in terms of most passes in the league. Note that our model considers the goal as a reward, but the reward was shared among agents and then the negative correlation with the goals was not strange. The results suggest that their ability indirectly contributing to the goal with many passes and other off-ball movements can be reflected in our indicator, rather than goals.

Next, we compared the contributions of on- and off-ball plays. In Fig. 3, there was a moderate positive correlation between average on- and off-ball Q-values ($\rho = 0.380$), suggesting that our indicator may tend to value similar play style between on- and off-ball situations. Because of a similar tendency, in the following analysis, we also used the Q-values including on- and off-ball cases.

Next, the relation between the average rating by the experts and the average Q-value is shown in Fig. 4a. There was no correlation between the two indicators ($\rho = -0.05$). This result may be related to the tendency that our indicator tended to value defense players who provided many passes, and there was a significant correlation between the goals and this subjective rating in a previous work [11].

Finally, we compared our results with existing off-ball indicators such as OBSO [10] and C-OBSO [11]. Figure 4b shows a low negative correlation with average OBSO ($\rho = -0.279$). Since OBSO values attacking players who will receive the ball, the tendency was similar to the correlation with the season goal. OBSO can also value the players' all time stamps, thus the offensive players can be valued separately based on both OBSO and our Q-values. For example, regarding Miyoshi (a midfielder) and Thiago (a defender), who had three and zero season goals, respectively, Miyoshi's performance can be highly valued in terms of off-ball movement to receive the ball (i.e., OBSO), and Thiago's can be highly valued in terms of passing and other off-ball movements (i.e., our method).

Figure 4c shows a low negative correlation with average C-OBSO ($\rho = -0.402, p > 0.05$). Compared with C-OBSO [11], which can value forward players such as Edigar and Nakagawa more than our Q-values, our Q-values tended to value midfielders and defenders such as Kida (0 season goals) and Hirose (1 season goal) indirectly contributing to the goals. From these results, we clarified the properties of our Q-values compared with those of conventional indicators.

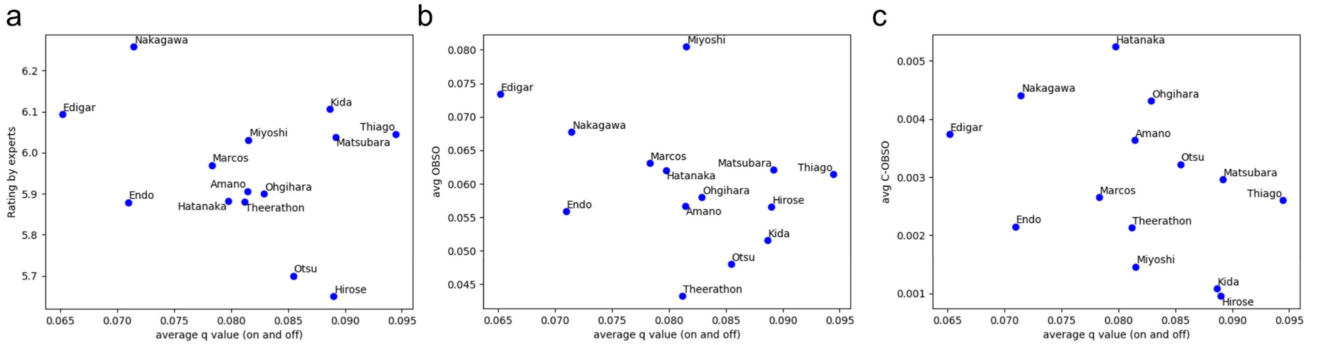


FIGURE 4. **The relationship between average Q-values by the proposed method and (a) the average rating by the experts, (b) OBSO [10], and (c) C-OBSO [11].**

V. CONCLUSION

In this paper, we proposed a comprehensive evaluation method for soccer players in attacking scenes using the framework of deep RL with action supervision. In the experiment, we analyzed the relationship with conventional indicators, season goals, and game ratings by experts, and showed the effectiveness of the proposed method. Our approach can assess how multiple players move continuously throughout the game, which is difficult to be discretized or labeled but vital for teamwork, scouting, and fan engagement.

A possible future direction is to improve RL models for a more valid valuation of players. The proposed RL model and Q-value computation can be further improved by more sophisticated rewards per action (e.g., [8]) and combining with a forward RL simulation e.g., ([39], [45]) while it is necessary to solve a gap between observation (data) and simulation spaces. Our approach can be also improved as offline RL (reviewed by [46]). For example, we did not consider the out of action distribution explicitly, which is a common problem in offline RL, and future work can tackle this problem.

ACKNOWLEDGMENT

The authors would like to thank Taiga Someya for his insightful code review, which contributed to the refinement of this project.

REFERENCES

[1] T. Decroos, L. Bransen, J. Van Haaren, and J. Davis, "Actions speak louder than goals: Valuing player actions in soccer," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 1851–1861.

[2] K. Toda, M. Teranishi, K. Kushiro, and K. Fujii, "Evaluation of soccer team defense based on prediction models of ball recovery and being attacked: A pilot study," *PLoS ONE*, vol. 17, no. 1, Jan. 2022, Art. no. e0263051.

[3] R. Umemoto, K. Tsutsui, and K. Fujii, "Location analysis of players in UEFA EURO 2020 and 2022 using generalized valuation of defense by estimating probabilities," 2022, *arXiv:2212.00021*.

[4] C. C. K. Yeung, T. Sit, and K. Fujii, "Transformer-based neural marked spatio temporal point process model for football match events analysis," 2023, *arXiv:2302.09276*.

[5] G. Liu and O. Schulte, "Deep reinforcement learning in ice hockey for context-aware player evaluation," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 3442–3448.

[6] G. Liu, Y. Luo, O. Schulte, and T. Kharrat, "Deep soccer analytics: Learning an action-value function for evaluating soccer players," *Data Mining Knowl. Discovery*, vol. 34, no. 5, pp. 1531–1559, Sep. 2020.

[7] K. Routley and O. Schulte, "A Markov game model for valuing player actions in ice hockey," in *Proc. 31st Conf. Uncertainty Artif. Intell. (UAI)*. Arlington, VA, USA: AUAI Press, 2015, pp. 782–791.

[8] P. Rahimian, J. Van Haaren, T. Abzhanova, and L. Toka, "Beyond action valuation: A deep reinforcement learning framework for optimizing player decisions in soccer," in *Proc. 16th Annu. MIT Sloan Sports Anal. Conf.*, Boston, MA, USA: MIT Press, 2022, p. 25.

[9] J. Fernández and L. Bornn, "Wide open spaces: A statistical technique for measuring space creation in professional soccer," in *Proc. 12th MIT Sloan Sports Anal. Conf.*, 2018, pp. 1–19.

[10] W. Spearman, "Beyond expected goals," in *Proc. 12th MIT Sloan Sports Anal. Conf.*, 2018, pp. 1–17.

[11] M. Teranishi, K. Tsutsui, K. Takeda, and K. Fujii, "Evaluation of creating scoring opportunities for teammates in soccer via trajectory prediction," in *Proc. Int. Workshop Mach. Learn. Data Mining Sports Anal.* Cham, Switzerland: Springer, 2022, pp. 53–73.

[12] K. Kurach, A. Raichuk, P. Stańczyk, M. Zając, O. Bachem, L. Espeholt, C. Riquelme, D. Vincent, M. Michalski, O. Bousquet, and S. Gelly, "Google research football: A novel reinforcement learning environment," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 4, pp. 4501–4510.

[13] K. Fujii, "Data-driven analysis for understanding team sports behaviors," *J. Robot. Mechtron.*, vol. 33, no. 3, pp. 505–514, Jun. 2021.

[14] A. Scott, K. Fujii, and M. Onishi, "How does AI play football? An analysis of RL and real-world football strategies," in *Proc. 14th Int. Conf. Agents Artif. Intell.*, vol. 1, 2022, pp. 42–52.

[15] P. Lucey, A. Bialkowski, M. Monfort, P. Carr, and I. Matthews, "Quality vs quantity: Improved shot prediction in soccer using strategic features from spatiotemporal data," in *Proc. MIT Sloan Sports Anal. Conf.*, 2014, pp. 1–9.

[16] T. Decroos, V. Dzyuba, J. Van Haaren, and J. Davis, "Predicting soccer highlights from spatio-temporal match event streams," in *Proc. AAAI Conf. Artif. Intell.*, vol. 31, no. 1, 2017, pp. 1302–1308.

[17] O. Schulte, M. Khademi, S. Gholami, Z. Zhao, M. Javan, and P. Desaulniers, "A Markov game model for valuing actions, locations, and team performance in ice hockey," *Data Mining Knowl. Discovery*, vol. 31, no. 6, pp. 1735–1757, Nov. 2017.

[18] L. Bransen and J. Van Haaren, "Measuring football players' on-the-ball contributions from passes during games," in *Proc. Int. Workshop Mach. Learn. Data Mining Sports Anal.* Cham, Switzerland: Springer, 2018, pp. 3–15.

[19] U. Dick, M. Tavakol, and U. Brefeld, "Rating player actions in soccer," *Frontiers Sports Act. Living*, vol. 3, p. 174, Jul. 2021.

[20] P. Power, H. Ruiz, X. Wei, and P. Lucey, "Not all passes are created equal: Objectively measuring the risk and reward of passes in soccer from tracking data," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2017, pp. 1605–1613.

[21] J. Brooks, M. Kerr, and J. Guttag, "Developing a data-driven player ranking in soccer using predictive model weights," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 49–55.

[22] U. Dick, D. Link, and U. Brefeld, "Who can receive the pass? A computational model for quantifying availability in soccer," *Data Mining Knowl. Discovery*, vol. 36, no. 3, pp. 987–1014, May 2022.

[23] D. Link, S. Lang, and P. Seidenschwarz, "Real time quantification of dangerousity in football using spatiotemporal tracking data," *PLoS ONE*, vol. 11, no. 12, Dec. 2016, Art. no. e0168768.

[24] K. Fujii, Y. Yoshihara, Y. Matsumoto, K. Tose, H. Takeuchi, M. Isobe, H. Mizuta, D. Maniwa, T. Okamura, T. Murai, Y. Kawahara, and H. Takahashi, "Cognition and interpersonal coordination of patients with schizophrenia who have sports habits," *PLoS ONE*, vol. 15, no. 11, Nov. 2020, Art. no. e0241863.

[25] P. Robberechts, "Valuing the art of pressing," in *Proc. StatsBomb Innov. Football Conf.* Bath, U.K.: StatsBomb, 2019, pp. 1–11.

[26] Y. Luo, O. Schulte, and P. Poupart, "Inverse reinforcement learning for team sports: Valuing actions and players," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 3356–3363.

[27] P. Rahimian and L. Toka, "Inferring the strategy of offensive and defensive play in soccer with inverse reinforcement learning," in *Proc. 8th Int. Workshop Mach. Learn. Data Mining Sports Anal. (MLSA)*. Cham, Switzerland: Springer, Sep. 2022, pp. 26–38.

[28] H. M. Le, Y. Yue, P. Carr, and P. Lucey, "Coordinated multi-agent imitation learning," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, vol. 70, 2017, pp. 1995–2003.

[29] H. M. Le, P. Carr, Y. Yue, and P. Lucey, "Data-driven ghosting using deep imitation learning," in *Proc. MIT Sloan Sports Anal. Conf.*, 2017, pp. 1–15.

[30] M. Teranishi, K. Fujii, and K. Takeda, "Trajectory prediction with imitation learning reflecting defensive evaluation in team sports," in *Proc. IEEE 9th Global Conf. Consum. Electron. (GCCE)*, Oct. 2020, pp. 124–125.

[31] K. Fujii, N. Takeishi, Y. Kawahara, and K. Takeda, "Policy learning with partial observation and mechanical constraints for multi-person modeling," 2020, *arXiv:2007.03155*.

[32] E. Zhan, S. Zheng, Y. Yue, L. Sha, and P. Lucey, "Generating multi-agent trajectories using programmatic weak supervision," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–14.

[33] R. A. Yeh, A. G. Schwing, J. Huang, and K. Murphy, "Diverse generation for multi-agent sports games," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4605–4614.

[34] L. Li, J. Yao, L. Wenliang, T. He, T. Xiao, J. Yan, D. Wipf, and Z. Zhang, "GRIN: Generative relation and intention network for multi-agent trajectory prediction," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 27107–27118, 2021.

[35] K. Fujii, K. Takeuchi, A. Kuribayashi, N. Takeishi, Y. Kawahara, and K. Takeda, "Estimating counterfactual treatment outcomes over time in complex multi-agent scenarios," 2022, *arXiv:2206.01900*.

[36] J. Fernández, L. Bornn, and D. Cervone, "Decomposing the immeasurable sport: A deep learning expected possession value framework for soccer," in *Proc. 13th MIT Sloan Sports Anal. Conf.*, 2019, pp. 1–20.

[37] N. Ding, K. Takeda, and K. Fujii, "Deep reinforcement learning in a racket sport for player evaluation with technical and tactical contexts," *IEEE Access*, vol. 10, pp. 54764–54772, 2022.

[38] T. Hester, M. Vecerik, O. Pietquin, M. Lanctot, T. Schaul, B. Piot, D. Horgan, J. Quan, A. Sendonaris, I. Osband, G. Dulac-Arnold, J. Agapiou, J. Leibo, and A. Gruslys, "Deep Q-learning from demonstrations," in *Proc. 32nd AAAI Conf. Artif. Intell., 30th Innov. Appl. Artif. Intell. Conf.*, 2018, pp. 3223–3230.

[39] K. Fujii, K. Tsutsui, A. Scott, H. Nakahara, N. Takeishi, and Y. Kawahara, "Adaptive action supervision in reinforcement learning from real-world multi-agent demonstrations," 2023, *arXiv:2305.13030*.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–15.

[41] JLEAGUE. (2019). *JLeague.jp 2019 Data*. [Online]. Available: https://www.jleague.jp/stats/2019/goal.html

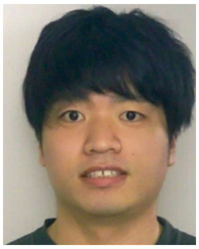
[42] Soccer-Digest. (2019). *Soccer Digest Web J1 Rating*. [Online]. Available: https://www.soccerdigestweb.com

[43] A. Kijima, K. Yokoyama, H. Shima, and Y. Yamamoto, "Emergence of self-similarity in football dynamics," *Eur. Phys. J. B*, vol. 87, no. 2, pp. 1–6, Feb. 2014.

[44] J. P. Guilford, *Fundamental Statistics in Psychology and Education*. New York, NY, USA: McGraw-Hill, 1950.

[45] T. Mendes-Neves, J. Mendes-Moreira, and R. J. Rossetti, "A data-driven simulator for assessing decision-making in soccer," in *Proc. 20th EPIA Conf. Artif. Intell. (EPIA)*. Cham, Switzerland: Springer, Sep. 2021, pp. 687–698.

[46] S. Levine, A. Kumar, G. Tucker, and J. Fu, "Offline reinforcement learning: Tutorial, review, and perspectives on open problems," 2020, *arXiv:2005.01643*.



HIROSHI NAKAHARA received the B.E. degree in electrical engineering and computer science from Kyushu University, Fukuoka, Japan, in 2020, and the M.S. degree from the Graduate School of Informatics, Nagoya University, Nagoya, Japan. His research interest includes data analysis in team sports, such as baseball and soccer.

KAZUSHI TSUTSUI received the B.A. degree from Tokyo Gakugei University and the M.S. and Ph.D. degrees from The University of Tokyo. After working on machine learning as a Researcher with Nagoya University, he joined the Faculty of Nagoya University, in 2020. He is currently a Designated Assistant Professor with the Graduate School of Informatics and also leads the Young Researcher Units for the Advancement of New and Undeveloped Fields, Nagoya University. His research interest includes multi-agent interactions, with applications extending to the study of team sports and animal group behaviors.

KAZUYA TAKEDA (Senior Member, IEEE) received the B.E., M.E., and Ph.D. degrees from Nagoya University. He was with Advanced Telecommunication Research Laboratories and KDD Research and Development Laboratories. In 1995, he joined Nagoya University, where he started a research group for signal processing applications. He is currently a Professor with the Graduate School of Informatics and the Green Mobility Collaborative Research Center, Nagoya University. His research interest includes behavior signal processing, including driving behavior.

KEISUKE FUJII (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Kyoto University, in 2009, 2011, and 2014, respectively. After his work as a Postdoctoral Fellow and a Research Scientist with Nagoya University and the RIKEN Center for Advanced Intelligence Project, Japan, he joined Nagoya University. He is currently an Associate Professor with the Graduate School of Informatics. His research interests include interdisciplinary studies among machine learning, behavioral sciences, and sports sciences.