Step-by-Step Acquisition of Cooperative Behavior in Soccer Task

Takashi Abe, Ryohei Orihara, Yuichi Sei, Yasuyuki Tahara, and Akihiko Ohsuga
Graduate School of Informatics and Engineering, The University of Electro-Communications, Tokyo, Japan
Email: abe.takashi@ohsuga.lab.uec.ac.jp, orihara@acm.org, {seiuny, tahara, ohsuga}@uec.ac.jp

***Abstract*—In this research, soccer task is investigated among the numerous tasks of deep reinforcement learning. The soccer task requires cooperative behavior. However, it is difficult for the agents to acquire the behavior, because a reward is sparsely given. Moreover, the behaviors of the allies and opponents must be considered by the agents. In addition, in the soccer task, if the agents attempt to acquire high-level cooperative behavior from low-level movements, such as ball kicking, a huge amount of time will be needed to learn a model. In this research, we conduct experiments in which reward shaping and curriculum learning are incorporated into deep reinforcement learning. This enables the agents to efficiently acquire cooperative behavior from low-level movements in a soccer task. The findings of this research indicate that reward shaping and curriculum learning with a designer's domain knowledge positively influence the agent's attempt to acquire cooperative behavior from low-level movements.**

***Index Terms*—soccer, multi-agent reinforcement learning, reward shaping, curriculum learning, MuJoCo**

I. Introduction

A. *Multi-agent Reinforcement Learning*

In recent years, deep reinforcement learning, which combines deep learning and reinforcement learning, has been widely used in game AI [1], [2] and robot control [3], [4]. However, its application in the real world still faces numerous challenges. In the real world, few tasks can be completed by a single agent, and cooperative behavior of multiple agents is required. Reinforcement learning that involves multiple agents is known as multi-agent reinforcement learning.

Multi-agent tasks are encountered in many real-world situations, such as peer-to-peer ridesharing system [5] and traffic signal control [6]. Li *et al.* addressed the order dispatching problem in the peer-to-peer ridesharing problem [5]. In the paper, they adopted the mean field approximation to simplify the local interactions, and their proposed method performed better than several strong baselines in the Accumulated Driver Income (ADI) and order response rate measures. Prabuchandran *et al.* applied multi-agent reinforcement learning algorithms to obtain dynamic traffic signal control policies, and modeled each traffic signal junction as an independent agent [6]. They showed through VISSIM that their algorithms performed better than the standard Fixed Signal Timing (FST) algorithm and the saturation balancing (SAT) algorithm at two real road networks. As described, multi-agent reinforcement learning is being studied to utilize reinforcement learning in the real world.

B. *Soccer Task*

Among the tasks of multi-agent reinforcement learning, a soccer task is most frequently used in literatures. In this task, elucidating which actions are likely to bring a reward and a punishment is difficult, as agents can only receive the reward or the punishment when they score a goal or they lose a point. In tasks in which the agents have few opportunities to receive rewards, learning is difficult [7]. Furthermore, in the soccer task, the agents find it difficult to score goals and prevent goals by themselves; thus, a cooperative and hostile behavior is necessary. Unlike learning by a single agent, the agents must select the optimal action, taking into consideration the behavior of the allies and opponents. Therefore, learning is difficult as the amount of information that needs to be considered increases, and the exploration takes a huge amount of time.

To date, a number of researches on reinforcement learning of a soccer task have been conducted. In particular, the Keepaway task [8] in RoboCup Soccer [9] is incorporated by numerous researchers. In this task, the agents are divided into the keepers team and the takers team. The keepers team holds the ball, so that the takers team cannot steal the ball. In most researches on soccer task, including the Keepaway task, agents learn to acquire a cooperative and hostile behavior through manually designed complex movements in advance. However, with regard to tasks with opponents, such as soccer simulations, human development efforts may be insufficient to outperform the opponent teams benefitting from the implementation by more knowledgeable designers [10]. This is because the designer's implementation ability is greatly reflected in the match results. Furthermore, if low-level movements, such as ball kicking and moving to the appropriate position, are designed in advance, the agents will be only able to acquire cooperative and hostile behavior based on the designer's perspective, thus making it difficult to obtain knowledge beyond the human knowledge.

In our research, we aim to improve learning efficiency by using reward shaping [11] and curriculum learning [12] in a soccer task. Reward Shaping [11] is a method that

Manuscript received October 28, 2021; revised February 10, 2022.

doi: 10.12720/jait.13.2.147-154

aims to improve learning efficiency by giving agents rewards step-by-step until they reach the final goal. In tasks in which agents sparsely receive rewards, there are few opportunities for agents to receive rewards until they reach the final goal, and it takes an enormous amount of time for them to learn behavior. Therefore, there has been a lot of research using reward shaping for soccer task in which it takes a lot of time for agents to learn the behavior. Curriculum Learning [12] defined by Bengio *et al.* is a method in which an agent first learns easy tasks and then gradually learns difficult tasks to solve a final goal. This research shows the effectiveness of curriculum learning in tasks such as classifying shapes and estimating the next word. When we learn something, we do not start with difficult tasks, but first learn easy tasks and then gradually learn difficult tasks. When agents learn, they can learn effectively by curriculum designed by us just like humans.

In this paper, we aimed to enhance learning efficiency using reward shaping [11] and curriculum learning to overcome a problem, in which a huge amount of time is needed for the agents to acquire cooperative behavior in soccer task. In addition, our research enables agents to acquire cooperative behavior by combining reward shaping with curriculum learning in a soccer task implemented with a physics engine MuJoCo [13] created by DeepMind. This paper is an extended version of [14].

II. Related Work

A. Soccer Task

So far, numerous researches on soccer task have been conducted, some of which enabled agents to acquire high-level behavior from low-level movements via reinforcement learning. Peng *et al.* [15] enabled agents to acquire dribbling skills by using hierarchical reinforcement learning, and Riedmiller *et al.* [10] conducted a research that allowed agents to acquire defensive behavior. However, when dribbling in an actual game, the agent has to carry the ball, taking into consideration the position of the allies and opponents. It is difficult for the agents to incorporate the dribbling skill learned in the environment without allies nor opponents into the cooperative behavior. In Riedmiller *et al.*'s research, in which agents acquired defensive behavior, there is a problem in the agent's ability to acquire cooperative behavior, taking into consideration the zone defense or the movements of the allies.

With the aim of solving these problems, Chitnis *et al.* investigated the agent's skills in passing the ball and receiving the pass from low-level movements [16]. They employed intrinsic motivation to encourage cooperative behavior. In this task, an episode ends when two agents are involved in a ball and then score a goal, or a certain amount of steps has passed. However, to acquire the cooperative behavior, it takes 100,000 episodes by using parallel learning of 30 agents. Furthermore, Liu *et al.* studied the reinforcement learning of a soccer task from low-level movements in a game format having two allies and two opponents [17]. In this research, the cooperative behavior of the agent is acquired using an algorithm combining SVG0 [18] and Population-Based Training [19]. In this task, an episode ends when either team scores a goal or 45 seconds have passed. However, to acquire pass skill or intercept skill, it takes 80 billion episodes in this learning. Therefore, since a huge amount of time is required to solve these problems, an appropriate approach for learning the task of acquiring cooperative behavior from low-level movements in soccer task needs to be established.

B. Reward Shaping

In tasks in which agents sparsely receive rewards, learning policy is difficult and very inefficient due to the difficulty in obtaining hints to achieve a final target during the agent's learning. For such tasks, reward shaping [11], in which agents gradually receive rewards, has been proposed. For example, consider the task of opening the lid of a plastic bottle. Simply twisting the lid will rotate the entire plastic bottle, and the lid cannot be opened. Learning would be relatively easy if the agent receives not only the extrinsic reward when the lid of the plastic bottle is opened but also the shaping reward when the plastic bottle is gripped with the opposite hand or the lid is twisted. However, learning will be affected if reward shaping is improperly set [11]. For example, when opening a plastic bottle, if the agents receive shaping rewards when the lid is twisted to the direction opposite to that in which it should be twisted, the wrong guidance will hinder an efficient learning. Therefore, the designer who has the domain knowledge of the task needs to decide how to give the agents the rewards. In our research, we use reward shaping to help agents acquire cooperative behavior efficiently in a soccer task, and we consider the effect of reward shaping on learning.

C. Curriculum Learning

There are some researches on the acquisition of behavior using curriculum learning in soccer task. The research of Narvekar *et al.* [20] focused on how to generate useful source tasks to be used in curriculum learning. They experimented with Half Field Offense task [21] in Robocup Soccer and showed that the proposed curriculum learning is efficient. In the research of Silva *et al.* [22], they generated tasks and curriculum automatically and showed learning efficiency in Half Field Offense task. However, in previous researches of curriculum learning in a soccer task [20], [22], agents cannot learn cooperative behavior. In the research of Silva *et al.*, an offensive agent learns behavior in an environment with one offensive agent and two defensive agents. In the research of Narvekar *et al.*, only an agent holding a ball learns behavior. Learning cooperative behavior of multiple agents is more difficult than that of a single agent without considering an ally's behavior, because the agents need to predict the ally's behavior in the latter case. The acquisition of cooperative behavior, which is a necessary skill in many situations, is an essential factor in learning in soccer task. However, researches have not been able to solve this problem so far. In our research, we focus on the acquisition of cooperative behavior by using curriculum learning in a soccer task and conduct our research in which agents learn cooperative behavior in a shooting chance.

III. Reinforcement Learning

Reinforcement learning is a learning method in which an agent observes the environment and maximizes the rewards for the actions the agent has taken. The total rewards that can be obtained by the agent in the future are defined as the value of the agent's action. In addition, the agent learns the policy, so that the policy maximizes the value of the agent's action. When the agent learns the policy, the method for the calculation of the value and selection of the policy that maximizes the value is called value-based method. Conversely, the method for deciding and improving the policy to enhance its value is called policy-based method.

A. Policy Gradient Method

To optimize the agent's action, a method can be used to identify the optimal policy called policy iteration. Policy iteration is a method that identifies the optimal policy by repeating the policy evaluation step of calculating the value function under a policy as well as the policy improvement step of updating the policy to enable maximization of the value function [23].

In a model-free environment, the policy gradient method is employed as an approach different from policy iteration. It is a policy-based method that updates θ in the direction of the gradient $\nabla_\theta J(\theta)$ of the policy parameter θ in the objective function $J(\theta)$ and improves it, so that the objective function $J(\theta)$ takes a larger value. The gradient $\nabla_\theta J(\theta)$ is presented below [24]:

$$\nabla_\theta J(\theta) = \mathbb{E}_\pi[\nabla_\theta log\pi(A_t|S_t,\theta)Q_\pi(S_t,A_t)] \tag{1}$$

here, $Q_\pi(S_t,A_t)$ denotes the value function when the agent takes action A_t in the state S_t at time t. Through this method, it becomes possible to improve the policy by incorporating the policy evaluation into the Q function.

B. REINFORCE

REINFORCE [25] is an algorithm that learns the policy through the approximation of the action-value function with the sum of the discounted rewards G_t in the policy gradient method. Suppose an agent takes action A_t in the state S_t at time t, R_{t+k} denotes the reward obtained in the future, and γ denotes the discount rate. The sum of the discounted rewards G_t and the gradient $\nabla_\theta J(\theta)$ with the policy parameter θ are presented below:

$$\mathrm{G}_t = \sum_{k=1}^{T-t}\gamma^{k-1}R_{t+k} \tag{2}$$

$$\nabla_\theta J(\theta) \approx \frac{1}{T}\sum_{t=0}^{T-1}\nabla_\theta log\pi(A_t|S_t)G_t \tag{3}$$

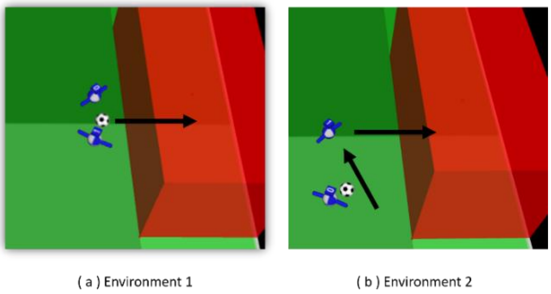
IV. Reinforcement Learning

A. Problem Setting

This paper incorporates reward shaping and curriculum learning into the REINFORCE algorithm to enable the agents to efficiently acquire cooperative behavior using MuJoCo Soccer [13].

1) Reward shaping

Two experiments in distinct problem settings have been conducted in this research, as presented in Fig. 1, to compare how learning proceeds in each case, with one involving the player's interaction and another without it. We conducted the first experiment with Environment 1 as a simple task, in which one of the two agents kicks a ball into the goal. In Environment 1, we consider that the final target is achieved when the agents score a goal, regardless of the movement of the ally. Environment 2 is similar that in the research by Chitnis *et al.* [16]. In this environment, we consider that the final target is achieved when two agents touch the ball and then score a goal. Even when humans play soccer, there are some situations, in which the probability of scoring a goal increases by selecting a pass in a shot chance. In Environment 2, we assume a situation in which such multiple agents cooperate, and evaluate whether reward shaping promotes multiple agents to take cooperative behavior.


(a) Environment 1

(b) Environment 2

Figure 1. Experimental environments.

2) Curriculum learning

Agents learn cooperative behavior in a shooting chance as shown in Fig. 2. Environment 3 consists of two offensive agents and two opponents. The offensive agents try to score a goal, while the defensive agents remain motionless as shown in Fig 2. In this environment, it is efficient to pass the ball to the ally in a distant position rather than shooting, and the offensive agents require cooperative behavior between agents. In actual games, there are many situations in which the agent chooses to pass rather than shoot, which is more likely to lead to a score. In Environment 3, the agents aim to acquire cooperative behavior in a shooting chance by using curriculum learning.

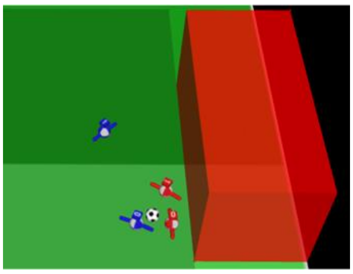


Figure 2. Experimental environments of environment 3.

B. Approach Overview

1) Learning algorithm

In this research, the model is trained using the algorithm proposed by Itoh *et al.* based on the REINFORCE algorithm [23]. In this model, the policy function was implemented as a Gaussian policy. Fig. 3 presents the whole picture of the learning algorithm in this research.

After the agents observe a state, they predict the optimum action using a stochastic policy based on the Gaussian model. Then, the observed state and reward are saved as history when they take an action. If the end condition is not fulfilled, the step will be continued. In this research, the agents received shaping reward, a reward that is given step by step until the final target is achieved, and extrinsic reward, which is a reward given when the final target is achieved. When an episode ends, the model is updated using states, actions, and rewards saved as a history for each step. To reduce the loss between the state value predicted by the agent based on the state at each step and the sum of the discounted rewards at each step, the state value function is updated. This function predicts the value of the state observed by the agent. In addition, the policy function is updated using the advantage function calculated based on the state value predicted by the state value function and the sum of the discounted reward. This function predicts the optimal action based on the state observed by the agent.

2) Reward shaping

In this research, reward shaping can facilitate the agent in acquiring cooperative behavior in a soccer task. In Environment 2, the agent will fail to receive the reward for achieving the final target, even if the agent scores a goal by itself. Therefore, the cooperative behavior of the two agents is necessary. In this environment, reward shaping is employed to enable agents to efficiently acquire useful behavior in this environment. However, as previously mentioned, due to the possibility that reward shaping negatively influences learning, how to give the agents rewards must be decided carefully. In this research, we give the agents extrinsic reward and shaping reward as follows, so that they can be actively involved in the ball.

Extrinsic Reward

- When two agents are involved in the ball and then score a goal: +10

Shaping Reward

- When the first agent is involved in the ball for the first time: +1
- When the second agent is involved in the ball for the first time: +1

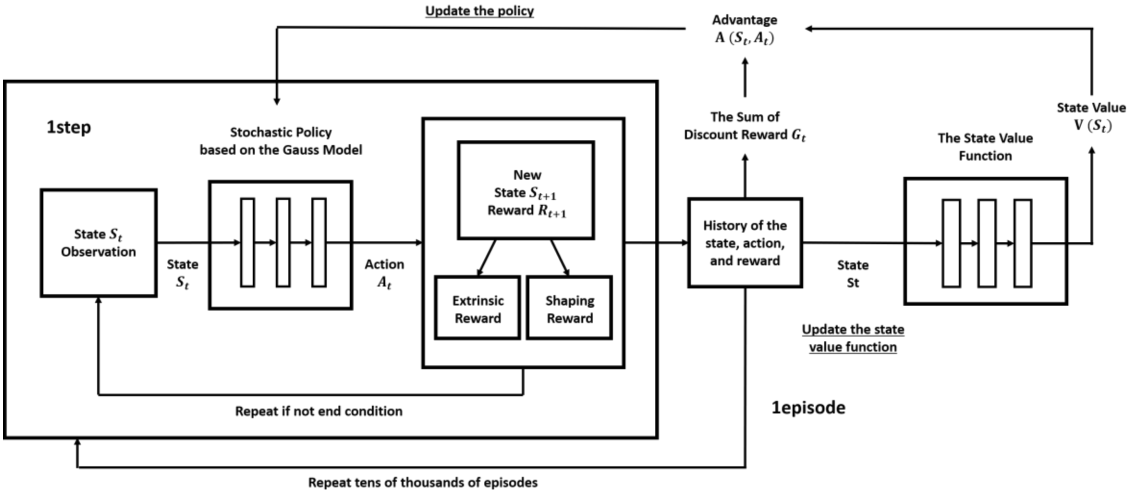


Figure 3. Learning algorithm in this research.

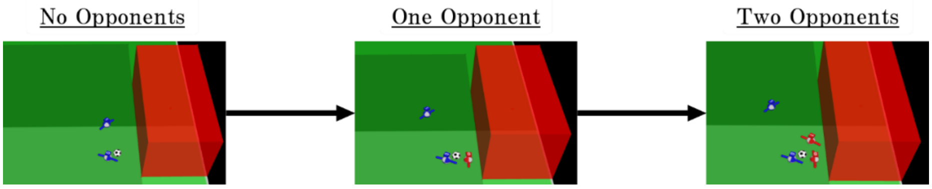


Figure 4. Overview of our proposed curriculum learning.

In this environment, due to the need for the agents to score a goal after the involvement of two agents in the ball, we need to make it easier for each agent to be involved in the ball. Thus, we give the agents rewards when each of them is involved in the ball. Here, the interaction of the two agents is encouraged by giving them rewards when the second agent is involved in the ball after another agent touched it, i.e., the second agent receives the pass. Furthermore, the shaping reward is set to be small and the extrinsic reward to be large to allow the agents to achieve the final target without being satisfied after they are involved in the ball. Moreover, we make it easier for the agents to achieve the final target.

3) Curriculum learning

In this paper, we use curriculum learning [12] to encourage agents to acquire cooperative behavior. When we practice soccer, we may first practice shooting with a cone as an opponent, and then practice shooting with a defender. This is because we can effectively acquire the behavior by practicing difficult exercises step by step, assuming a certain scene. Following this practice, agents first learns to pass a ball without any defensive agents.

Then, we prepare some environments with more defensive agents, and agents learn cooperative behavior step by step. An overview of curriculum learning is shown in Fig. 4. As shown in Fig. 4, we prepare three environments. First, agents try to score a goal with no opponents. Second, the agents try to score a goal with one opponent. Third, the agents try to score a goal with two opponents. Agents learn cooperative behavior in the first environment. Next, the agents learn cooperative behavior in the second environment using the model in which agents trained in the first environment. Finally, the agents learn cooperative behavior in the third environment using the model in which the agents trained in the second environment. In this experiment, we set a situation that the final target is achieved if two agents are involved in the ball and score a goal. Even if only one agent scores a goal, the final target is not achieved. Therefore, the agents must select cooperative behavior in this environment. We conduct experiments on agent's acquisition of cooperative behavior in this environment and verify the effect of our proposed curriculum learning in a soccer task.

V. Experiments

A. Experiment Summary

In this research, a soccer task is simulated in MuJoCo Soccer, which is a physics engine developed by DeepMind. The soccer task is simulated under the following conditions [13].

- *State*: The state space has a dimension of 81. The dimensions include the information of the proprioception, such as position, velocity, and accelerometer; the situation at a scene, such as an egocentric ball position, velocity and angular velocity, goal and corner positions; and teammate and opponent, such as orientation, position, and velocity.
- *Action*: The action space has a dimension of three. It includes accelerating the body forward-backward, rotating the body, and applying downward force to jump.
- *Episode*: One episode finishes when the final target is achieved or after 10s.

With regard to the policy, the input is the 81-dimensional state, and the output is the 3-dimensional action. The policy is implemented using a three-layer fully connected neural network with 810, 220, and 60-unit hidden layers, respectively. With regard to the value function model, the input is the 81-dimensional state, and the output is the 1-dimensional action value. The value function is implemented using a three-layer fully connected neural network with 810, 63, and 5-unit hidden layers, respectively. The learning rate is 0.99. RMSProp is employed in learning the policy as the optimization algorithm, and Adam is used in learning the value function model.

B. Results

1) Environment 1

For Environment 1, in which the agents are unnecessary to interact with each other, learning efficiency is evaluated based on the average target achievement rates. The results are presented in Fig. 5. In this task, the agents only receive the extrinsic reward when they achieve the final target, and learning is easily stabilized in an early stage. Even if the agents acquire a high-level behavior from low-level movements, it will be relatively easy for them to learn the high-level behavior without cooperative behavior. This is due to the relatively narrow state space which the agents consider comparing with Environment 2.

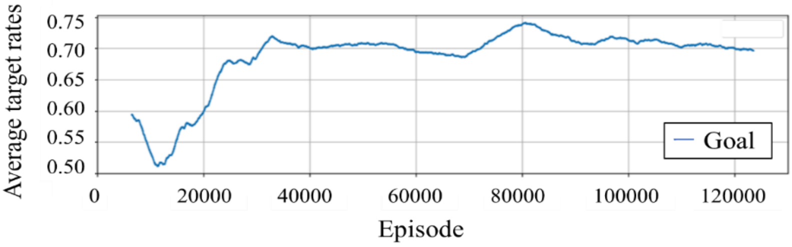
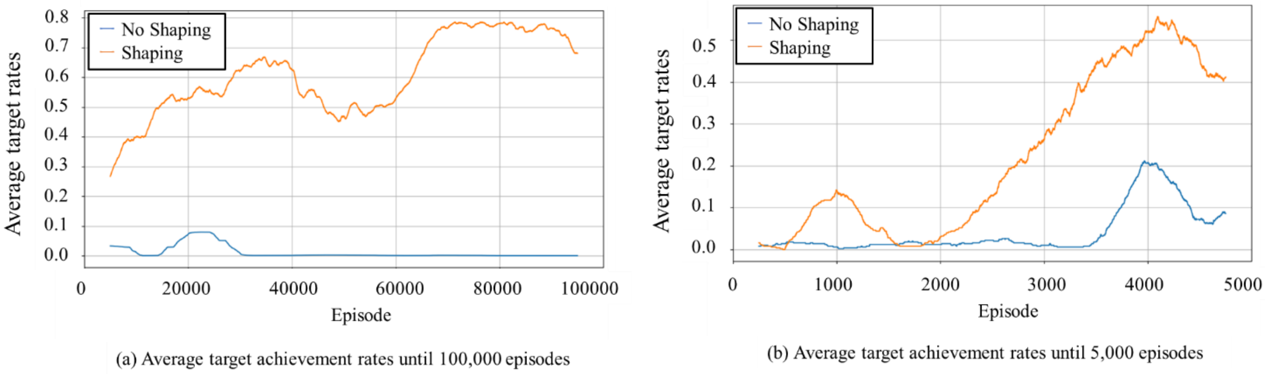


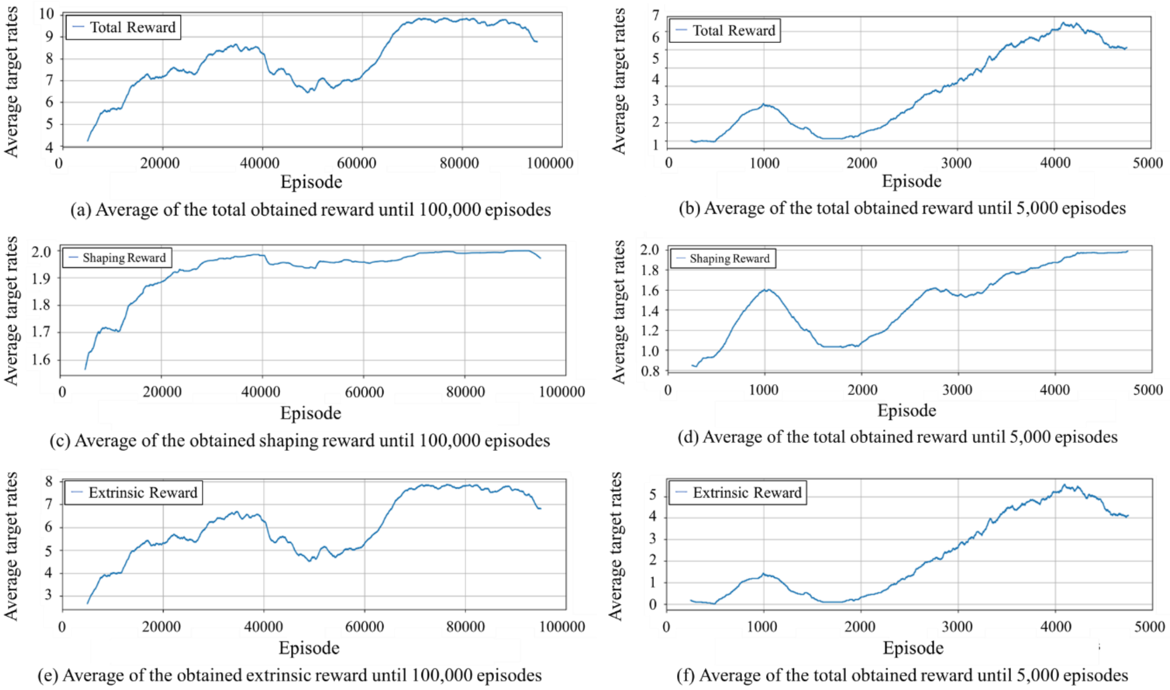
Figure 5. Average target achievement rates in Environment 1.



(a) Average target achievement rates until 100,000 episodes

(b) Average target achievement rates until 5,000 episodes

Figure 6. Average target achievement rates in Environment 2.



(a) Average of the total obtained reward until 100,000 episodes

(b) Average of the total obtained reward until 5,000 episodes

(c) Average of the obtained shaping reward until 100,000 episodes

(d) Average of the total obtained reward until 5,000 episodes

(e) Average of the obtained extrinsic reward until 100,000 episodes

(f) Average of the total obtained reward until 5,000 episodes

Figure 7. Average of the obtained reward in the case of using reward shaping.

2) Environment 2

For Environment 2, in which the agents need to interact with each other, learning efficiency is evaluated based on the average target achievement rates. Fig. 6 presents the results from the two learning conditions, one with reward shaping and another without it. As a result of the learning, the average target achievement rate in the case with reward shaping is significantly higher than that without it.

In the early stage of learning up to 5000 episodes, target achievement rates with reward shaping rapidly increase from about 2000 episodes. Moreover, target achievement rates without reward shaping begins to increase for the first time in 3500 episodes. These experimental results indicate that reward shaping promotes the cooperative action of the two agents.

The experimental results without reward shaping indicate that acquiring the behavior from low-level movements is more difficult when agents need to interact than when they do not. Contrary to the case in which complex movements are implemented in advance, exploration is difficult due to many choices of the agent's actions when learning the behavior from low-level movements. In this environment, an agent who passes the ball must grasp the position of the ally and make a pass toward there; then, an agent who receives the ball must learn how to control the ball and score a goal. This series of actions need to be learned by two agents. Since the agents will find it difficult to learn if a reward is given only when all of these actions are completed, the use of reward shaping is suggested for this task to increase the number of times when the rewards are given.

Fig. 7 presents the average of the total obtained reward, the average of the obtained shaping reward, and the average of the obtained extrinsic reward when using reward shaping in learning. Looking at the average of the obtained shaping reward, it can be determined that the reward finally converges to 2.0, and the two agents seem to learn to be involved in the ball. In the early stage of learning, it is difficult for the agents to obtain 1.0 shaping reward. Moreover, there are episodes in which even one agent cannot be involved in the ball, but two agents can be involved in the ball after 5000 episodes. From this, it can be inferred that reward shaping positively influences learning in the early stage.

Furthermore, the obtained shaping reward has increased without falling below 1.0 from 2,000 episodes. This resulted in the sharp increase of the obtained extrinsic reward from 2,000 episodes. From this observation, it can be inferred that if two agents are easily involved in the ball at the beginning, that is, the first agent finds it easier to kick the ball toward the second agent and increase the chances of the shoot, a higher target achievement rate can be obtained.

3) Environment 3

We experiment in the environment shown in Fig. 2, using schemes named Curriculum and No Curriculum. Curriculum denotes that agents learn cooperative behavior using our proposed curriculum learning. No Curriculum denotes that the agents learn cooperative behavior without prior learning. Our proposed curriculum consists of a task with no opponents and a task with an opponent, and the offensive agents learn 10000 episodes in each task. Fig. 8 shows the average target achievement rate.

The average target achievement rate for Curriculum is higher than that for No Curriculum. Up to 1000 episodes, the target achievement rate is about 10% in both approaches. However, after that, Curriculum learning accelerates. In particular, after 2500 episodes, the agents using curriculum learning achieve the target nearly 60% of the time. By learning the task in which there are no opponents and the task in which there is an opponent, the agents can acquire the skill of passing a ball and achieve the target faster even when there are more opponents. On the other hand, in No Curriculum, the target achievement rate remains low. Even if the agents receive the shaping rewards, the agents have little chance to receive the rewards if there are some obstacles, and it is difficult for agents to acquire cooperative behavior.

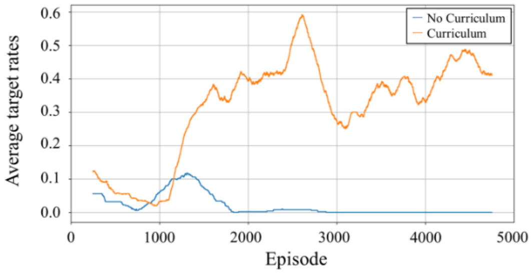


Figure 8. Average target achievement rates in Environment 3.

VI. Conclusion

In this paper, a learning method that uses reward shaping and curriculum learning in a soccer task, in which the agents must acquire cooperative behavior, is proposed. In this research, we show that the average target achievement rate in the case with our proposed method is significantly higher than that without it in the soccer task, and confirmed that reward shaping and curriculum learning can effectively facilitate the agent in acquiring cooperative behavior. In the future, how to promote learning efficiency by enabling agents to learn in different environments in stages will be studied.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Takashi Abe conducted the research and wrote the paper. All authors except the first author corrected this paper. All authors had approved the final version.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Numbers JP17H04705, JP18H03229, JP18H03340, JP18K19835, JP19H04113, JP19K12107.

References

[1] V. Mnih, *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529-533, February 2015.

[2] D. Silver, *et al.*, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, pp. 484-489, January 2016.

[3] T. P. Lillicrap, *et al.*, "Continuous control with deep reinforcement learning," in *Proc. 4th International Conference on Learning Representations*, 2015.

[4] D. Silver, *et al.*, "Deterministic policy gradient algorithms," in *Proc. the 31st International Conference on International Conference on Machine Learning*, 2014, pp. 387-395.

[5] M. Li, *et al.*, "Efficient ridesharing order dispatching with mean field multi-agent reinforcement learning," in *Proc. The World Wide Web Conference*, 2019, pp. 983-994.

[6] K. J. Prabuchandran, A. N. H. Kumar, and S. Bhatnagar, "Multi-agent reinforcement learning for traffic signal control," in *Proc. 17th International IEEE Conference on Intelligent Transportation Systems*, 2014, pp. 2529-2534.

[7] R. S. Sutton and A. G. Barto, *Introduction to Reinforcement Learning*, Cambridge, MA, USA: MIT Press, 1998.

[8] P. Stone, R. S. Sutton, and G. Kuhlmann, "Reinforcement learning for robocup soccer keepaway," *Adaptive Behavior*, vol. 13, no. 3, pp. 165-188, September 2005.

[9] H. Kitano, M. Asada, Y. Kuniyoshi, I. Noda, and E. Osawa, "Robocup: The robot world cup initiative," in *Proc. the First International Conference on Autonomous Agents*, 1997, pp. 340-347.

[10] M. Riedmiller, T. Gabel, R. Hafner, and S. Lange, "Reinforcement learning for robot soccer," *Autonomous Robots*, vol. 27, no. 1, pp. 55-73, July 2009.

[11] A. Y. Ng, D. Harada, and S. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping," in *Proc. the Sixteenth International Conference on Machine Learning*, 1999, pp. 278-287.

[12] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. the 26th Annual International Conference on Machine Learning*, 2009, pp. 41-48.

[13] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 5026-5033.

[14] T. Abe, R. Orihara, Y. Sei, Y. Tahara, and A. Ohsuga, "Acquisition of cooperative behavior in a soccer task using reward shaping," in *Proc. 5th International Conference on Innovation in Artificial Intelligence*, 2021.

[15] X. B. Peng, G. Berseth, K. Yin, and M. v. d. Panne, "DeepLoco: Dynamic locomotion skills using hierarchical deep reinforcement learning," *ACM Trans. Graph.*, vol. 36, no. 4, article 41, July 2017.

[16] R. Chitnis, S. Tulsiani, S. Gupta, and A. Gupta, "Intrinsic motivation for encouraging synergistic behavior," in *Proc. 8th International Conference on Learning Representations*, 2020.

[17] S. Liu, G. Lever, J. Merel, S. Tunyasuvunakool, N. Heess, and T. Graepel, "Emergent coordination through competition," in *Proc. 7th International Conference on Learning Representations*, 2019.

[18] N. Heess, G. Wayne, D. Silver, T. Lillicrap, T. Erez, Y. Tassa, "Learning continuous control policies by stochastic value gradients," in *Proc. the 28th International Conference on Neural Information Processing Systems*, 2015, vol. 2, pp. 2944-2952.

[19] M. Jaderberg, *et al.*, "Population based training of neural networks," arXiv:1711.09846, 2017.

[20] S. Narvekar, J. Sinapov, M. Leonetti, and P. Stone, "Source task creation for curriculum learning," in *Proc. the 15th International Conference on Autonomous Agents and Multiagent Systems*, 2016, pp. 566-574.

[21] S. Kalyanakrishnan, Y. Liu, and P. Stone, "Half field offense in RoboCup soccer: A multiagent reinforcement learning case study," in *Proc. Robot Soccer World Cup X*, 2006, pp. 72-85.

[22] F. L. D.Silva and A. H. R. Costa, "Object-Oriented curriculum generation for reinforcement learning," in *Proc. the 17th International Conference on Autonomous Agents and MultiAgent Systems*, 2018, pp. 1026-1034.

[23] T. Itoh, *et al.*, "You can use this in your field! Introduction to Deep Reinforcement Learning using Python, Exploration and Control by Reinforcement Learning," Shoeisha, 2019. (in Japanese)

[24] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Proc. the 12th International Conference on Neural Information Processing Systems*, 1999, pp. 1057-1063.

[25] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Mach. Learn.*, vol. 8, no. 3, pp. 229-256, May 1992.

Takashi Abe received BE degree in information science and engineering from the University of Electro Communications (UEC), in 2019. In 2021, he received ME degree in information science and engineering from The University of Electro Communications (UEC). His research interests include artificial intelligence. He has been with FUJIFILM Business Innovation Corporation, Tokyo, Japan, since 2021.

Ryohei Orihara received his Ph.D. degree in engineering from the University of Tsukuba in 1999. He joined Toshiba Corporation, Kawasaki, Japan, in 1988. He has been with Kioxia Corporation, Tokyo, Japan, formerly known as Toshiba Memory Corporation, since 2019, where he is currently a Senior Fellow at the Digital Transformation Technology R&D Center. He has also been a Guest Professor at the University of Electro-Communications, Tokyo, Japan, since 2010. His current research interests include artificial intelligence, machine learning, data mining and text mining. He served as a vice president of JSAI during 2017-2019. He is an IPSJ fellow.

Yuichi Sei received the Ph.D. degree in information science and technology from the University of Tokyo in 2009. From 2009 to 2012, he was with the Mitsubishi Research Institute. He joined the University of Electro-Communications in 2013, and is currently an associate professor in the Graduate School of Informatics and Engineering. He is also a visiting researcher at Mitsubishi Research Institute and an adjunct researcher at Waseda University. His current research interests include pervasive computing, privacy-preserving data mining, and software engineering.

Yasuyuki Tahara is an associate professor in the University of Electro-Communications. He received his BSc and his MSc in Mathematics from the University of Tokyo, Japan, and his PhD in Information and Computer Science from Waseda University, Japan, in 1989, 1991, and 2003, respectively. He joined Toshiba Corporation in 1991. He was a visiting researcher in City University London, UK, from 1995 to 1996, and in Imperial College London, UK, from 1996 to 1997. He left Toshiba Corporation and joined NII in 2003. He left NII and joined the University of Electro-Communications in 2008. His research interests include formal verification of software and requirements engineering. Prof. Tahara is a member of the Information Processing Society of Japan, the Institute of Electrical Engineers of Japan, and Japan Society for Software Science and Technology.

Akihiko Ohsuga received his Ph.D. degree in computer science from Waseda University in 1995. From 1981 to 2007 he was with Toshiba Corporation. He joined the University of Electro-Communications in 2007. He is currently a professor in the Graduate School of Informatics and Engineering. He is also a visiting professor at National Institute of Informatics. His research interests include agent technologies, web intelligence, and software engineering. He is a member of IEEE Computer Society (IEEE CS), Information Processing Society of Japan (IPSJ), Institute of Electronics, Information and Communication Engineers (IEICE), Japanese Society for Artificial Intelligence (JSAI), Japan Society for Software Science and Technology (JSSST), and Institute of Electrical Engineers of Japan (IEEJ). He has been a fellow of IPSJ since 2017. He served as a Chair of IEEE CS Japan Chapter, a member of JSAI Board of Directors, a member of JSSST Board of Directors, and a member of JSSST Councilor. He received IPSJ Best Paper Awards in 1987 and 2017.