Multi-Agent Visual Coordination Using Optical Wireless Communication

Haruyuki Nakagawa and Asako Kanezaki, *Member, IEEE*

***Abstract*—Communication is a key element in applying multi-agent reinforcement learning to a wide range of real-world scenarios. We focus on optical wireless communication (OWC), which is a practical solution to be used in various real situations where radio communication is not available, such as underwater or in a lot of radio noise environment. OWC is a method of communicating only with other agents in visual range using light, unlike radio wave like communication which is mostly assumed in existing research on multi-agent reinforcement learning. Due to limited communication, when OWC is used, overall performance is generally degraded from the case with full communication. In this letter, we propose a reinforcement learning method that learns visual coordination behavior using OWC. Our proposed visually cooperative behavior enables agents equipped with limited field of view (FOV) cameras to efficiently comprehend and imagine their surrounding environment through cooperative communication. Experimental results in simulation demonstrated that, using the proposed visual coordination method, multi-agents using OWC with general FOV show comparable performance to those with radio wave like full communication. Additionally, it has been demonstrated that this method can improve performance in various multi-agent reinforcement learning algorithms. We also implement OWC devices on real mobile robots and demonstrated the proposed multi-agent operation.**

***Index Terms*—Multi-robot systems, cooperating robots.**

I. Introduction

A MULTI-AGENT system is more complex than a single-agent system because of the difficulty of planning the interaction of multiple agents. Multi-agent reinforcement learning (MARL) is thus helpful for the system in which each agent learns to cooperate along with self-motion [15], [17]. Existing letters on MARL mostly assume that agents can communicate with all other agents. In other words, they assume the full communication scenarios such as using radio communication, which is possible regardless of the location of the agents and obstacles in the environment. However, in general, there are many situations where full communication is impossible in the real world. For example, radio communication is difficult in underwater environments because radio waves are absorbed and attenuated. It is known that robots that explore the sea are therefore connected by wired lines for communication, and thus the coordinated actions of multiple agents are limited [16].

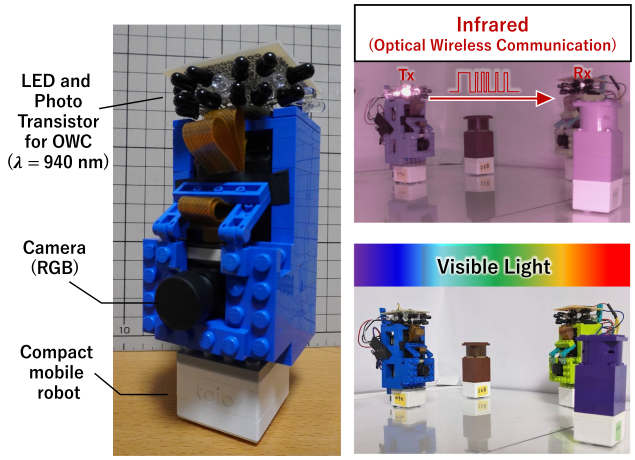


Fig. 1. Physical agents used in the MARL system with optical wireless communication (OWC) in this letter.

Recently, camera-based limited visual input is utilized as the most common way for agents to obtain information in real-world multi-agent systems [8], [24]. Reinforcement learning using limited field of view (FOV) is a partially observed Markov decision process, which is more difficult to learn than the case with fully observed states. Meanwhile, one advantage of a multi-agent system is that it is possible for each agent to obtain additional information beyond its own FOV by communicating with each other. However, it is impractical to share all the observation information obtained by other agents due to the limited communication bandwidth and response time.

In this letter, we examine the issues that arise in real-world applications of MARL from the viewpoints of communication methods and visual information sharing. In particular, we investigate optical wireless communication (OWC) for our MARL system. OWC is applicable to challenging environments such as underwater [18] and space [11], as well as for recent applications such as communication between autonomous vehicles and traffic signals [20]. OWC is also suitable for multi-agent systems such as swarm robotics because it can be implemented with only low-cost devices. As shown in Figs. 1 and 2, we use optical devices such as LEDs, photodiodes, and cameras for OWC. As

Manuscript received 31 March 2023; accepted 21 July 2023. Date of publication 14 August 2023; date of current version 17 October 2023. This letter was recommended for publication by Associate Editor G. Pereira and Editor M. Ani Hsieh upon evaluation of the reviewers' comments. This work was supported by JST FOREST Program, Japan under Grant JPMJFR206H. *(Corresponding author: Asako Kanezaki.)*

The authors are with the Department of Computer Scienceng, Tokyo Institute of Technology, Tokyo 152-8550, Japan (e-mail: nakagawa.h.aj@m.titech.ac.jp; kanezaki@c.titech.ac.jp).

This letter has supplementary downloadable material available at https://doi.org/10.1109/LRA.2023.3304905, provided by the authors.

Digital Object Identifier 10.1109/LRA.2023.3304905

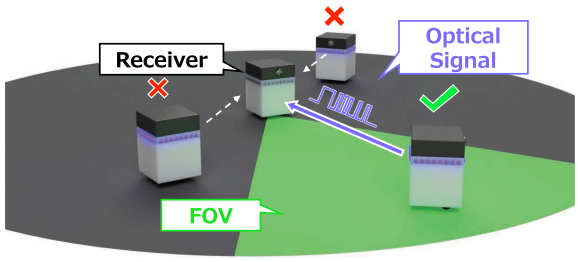


Fig. 2. Conceptual diagram of OWC. Each agent has a limited field of view as a receiver and can only receive signals from agents within its field of view (FOV).

shown in this figure, the assumption of FOV in the receiver side of the OWC is the same as vision, while the LEDs in the transmitter side can be recognized from all 360° directions.

Furthermore, we propose a visual coordination learning method for MARL. Unlike ordinary wireless communication, OWC is characterized by the fact that the communication partner must be inside its field of view. Therefore, control of visual direction is important to increase performance. In our method, agents have two networks, one of which predicts what they will see when they turn to look in different directions and the other determines which direction they are looking in is most likely to yield the next highest reward. In this way, the proposed method allows for actions that maximize the benefits of communication with other agents.

The contributions of this letter are as follows.

- We propose a method for controlling visual direction by sharing visual information with multiple agents and predicting the best direction to view.
- We show that the proposed visual coordination method reduces the performance gap between OWC and full communication in various multi-agent tasks, while also improving the overall success rate of these tasks.
- We implement OWC with the proposed method on a real robot system and confirmed its effectiveness in multi-agent cooperative tasks.

II. Related Work

Multi-agent system with communication: In a multi-agent environment, there have been many reports examining the effects of using communication between agents to enhance the task completion while promoting behavioral diversity. Forester et al. [6] introduced Differentiable Inter-Agent Learning (DIAL) in a deep recurrent Q network and showed high performance for DIAL by using a differentiable end-to-end communication channel. And, Lowe et al. [15] proposed Multi-Agent Deep Deterministic Policy Gradient (MADDPG) algorithm and demonstrated that a high level of coordinated behavior can be learned even in adversarial tasks and with communication. They also employ deterministic policy gradient methods to handle continuous actions. In addition, several articles have reported on the demonstration of communication capabilities in mobile vehicles, taking into account real-world constraints such as communication distance, errors, and budget [1], [9], [22].

Optical wireless communication: The characteristics of OWC have been extensively described as a method of communication between underwater access points and submarines in environments where radio communication is difficult [18]. As in these reports, communication functions are possible with inexpensive photodiodes and LEDs, and are easy to implement for small agents. In recent years, there have been reports of OWC and visible light communications (VLC) using cameras with two-dimensional arrays of light-receiving elements instead of point data [3], [23], [25]. Yamazato et al. [25] have studied the possibility of equipping vehicles with cameras and communicating with LEDs, such as traffic lights. In addition, an event-based camera that detects only events in which the brightness value changes, has been developed in recent years [4], [7]. Event-based cameras are recently used for VLC because they can operate at a higher frame rate than a conventional camera [3], [23]. It is expected that OWC will expand to robotics for various environments in the future in terms of acquiring communication information using visual information at high frame rates. To our knowledge, no research has been conducted on reinforcement learning utilizing OWC in multi-agent mobility devices.

Coordination of limited visual information: There are not many studies that aim to achieve collaboration by utilizing communication between multiple agents with limited FOV. Baker et al. [2] showed that agents with a limited 135° FOV learn how to use tools through cooperative and adversarial behavior. However, these agents are also assumed to have lidar-like all-surrounding sensors, so they are not completely limited vision. In terms of multiple agents coordinating their vision, several methods have been proposed [13], [14], [21]. Wu et al. [24] proposed a method that incorporates the positions of other agents and self subjective image information to generate an overhead map. This method uses the generated map as input and the agent's movements as actions to learn the distribution of Q values through a network and make decisions. This method requires RGB-Depth sensor to obtain subjective images to generate the overhead map. Generating an overhead map and the approach of creating and maintaining maps using Simultaneous Localization and Mapping (SLAM) may be computationally demanding and consume a large amount of memory. In contrast, we consider a visual cooperation algorithm using merely an RGB sensor that can be easily equipped on inexpensive multi-robots.

III. Method

A. Multi-Agent Tasks

In this letter, we validate the effectiveness of the proposed method on several well-known multi-agent tasks: *Simple spread*, *Predator-Prey*, and *Keep-Away*. For the sake of clarity, we hereby describe the experimental setup using the *Simple Spread* task. The N agents $\{a_1, \ldots, a_N\}$ and N landmarks $\{l_1, \ldots, l_N\}$ are used in the task and each agent has to reach a different landmark. Basic information regarding *Simple Spread* task is provided below.

- *Action space:* The visual direction of the next step θ $(-\pi \leq \theta \leq \pi)$ and the amount of movement of the agent in that direction s $(0 \leq s \leq S)$, which is the action of one step.

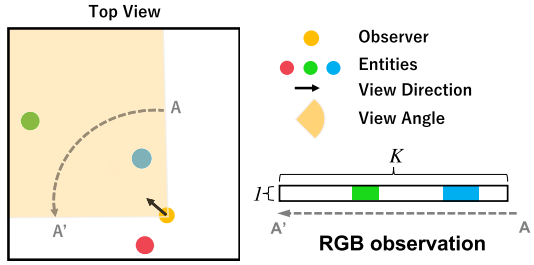


Fig. 3. 1D-RGB observation in our experiments.

Note that S is the parameter of the maximum distance that can be moved in single step.

- *Observation information:* 1-dimensional visual RGB information that scans a range of viewing angles centered on its own viewing direction (Fig. 3) and individual state information such as position $\mathbf{p}_i$ and visual direction $\mathbf{d}_i$.
- *Reward:* For *Simple Spread* task, reward r is given as penalty by taking the sum of distances to the nearest agent for each landmark position $\mathbf{q}_i$. This is the same as used in existing letters [15].

$$r = -\sum_{i=1}^{N} \min_{j=1,\ldots,N} \|\mathbf{q}_i - \mathbf{p}_j\|. \tag{1}$$

- *Communication* : The communication signal consists of position $\mathbf{p}_i$, view direction $\mathbf{d}_i$, and topological information (to be described in III-C).

Note that agent is not aware of the location of landmarks.

B. OWC for MARL

In OWC, the transmitter sends a blinking pattern of light from an LED or other device as a signal, and the receiver decodes it with a photodetector or other device. Due to the straightness of this light, OWC requires that the transmitting agent (Tx) be visible to the receiving agent (Rx) due to its communication mechanism. If a_j (Tx) is not visible to a_i (Rx), a_i will not receive a signal from a_j. Also, it cannot obtain a signal even if there is some other object between Rx and Tx. In this experimental system, there are no restrictions on the distance that can be communicated by OWC.

Define a variable v_{im} that represents whether an entity e_m $(1 \leq m \leq M)$ in the visual range. Entity e_m represents agent a_m or landmark l_m, and the total number of these $M = 2\,N$ in *Simple Spread* task. If the mth entity is within the visual range of the ith agent, $v_{im} = 1$, and if it is not within the visual range, $v_{im} = 0$. Yali et al. [5] define the graph relationship with communicating agents as topology. In this letter, we refer to such "what each agent is looking at" as *topological information*. Agent a_i extracts and maintains $\{v_{im}\}_{m=1}^{M} \in \mathbb{R}^M$ (which we call self topological information) from its own visual information. In addition to the location information $\mathbf{p}_i$ and visual direction $\mathbf{d}_i$, it is shared with other agents. By adding topological information received from other agent a_m via communication, each agent a_i acquires $V_i = (v_{nm}) \in \mathbb{R}^{N \times M}$. Since this information is binary, the bandwidth required for communication is small and considered realistic. An example of topological information V_i is shown in Fig. 4. As shown in Fig. 4(a), in this example, there is a green agent in the red agent's field of view and a black landmark in the green agent's field of view. The other two landmarks are in the field of view of the blue agent. A topological graph representing these situations is shown in Fig. 4(b).

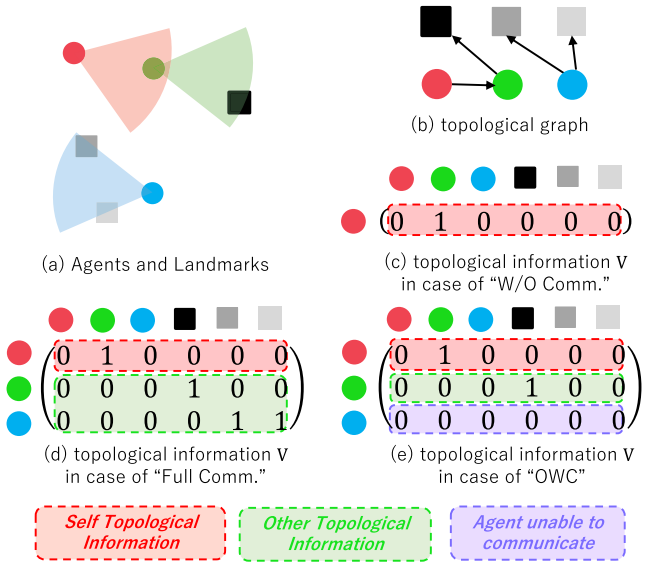


Fig. 4. Exemplar topological information V of Red agent in different communication methods. Circles represent agents and rectangles represent landmarks.

For each communication method, topological information V_i for red agent is shown in Fig. 4(c), (d), and (e). In a case of w/o communication, only the self-topology information of each agent is utilized as V_i (Fig. 4(c)). In a case of full communication, all agents can obtain topology information from each other through communication, resulting in $V_1 = V_2 = \ldots = V_N$ (Fig. 4(d)). However, it should be noted that when in a case of OWC, topological information of agents that are not within the field of view (in this case, the blue agent) is not available (Fig. 4(e)). In such cases, the corresponding row values of these agents are filled with 0.

C. How to Control Visual Direction

The proposed algorithm for selecting the visual direction is shown in Fig. 5. The basic idea of this method is to share "what is seen" among agents as compressed information that can be communicated to each agent, so that they can imagine areas that are not seen. Then, information such as what the agents are seeing is used to determine whether a high reward can be obtained. A simple way to control visual direction among multiple agents is to have all agents actually look in various directions, such as through random sampling or grid search. However, this method requires a large number of steps, and the number of steps increases exponentially when considering combinations of multiple agents, which worsens sample efficiency. Another possibility is to reward the control of visual direction itself, but it is difficult to set the optimal balance between visual and

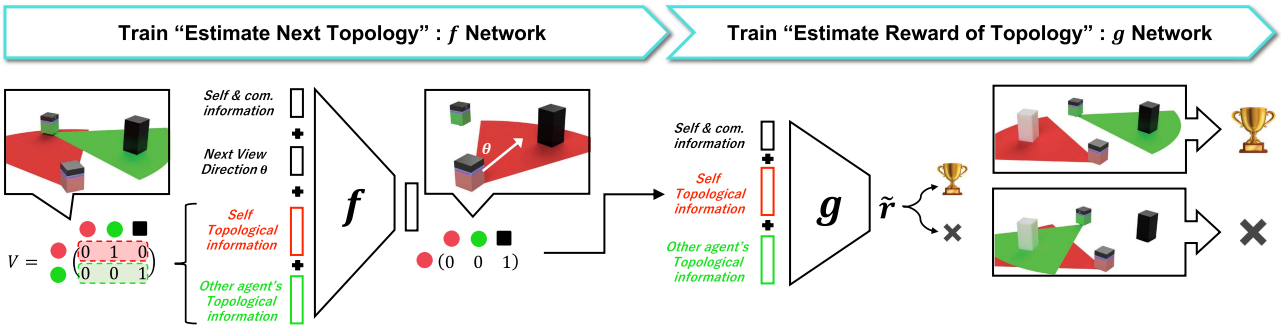


Fig. 5. Diagram of visual coordination methods among multiple agents.

behavioral rewards in the task with actions being considered in this study. In contrast, the proposed method prevents an increase in the number of steps by imagining what is visible according to the visual direction, and applies a reward setting that leads to task success by linking the visual direction control method to the reward setting by the action. Moreover, by using only abstract information without relying on detailed visual input, it becomes easier to imagine the situation when changing visual direction regardless of the actual circumstances.

The proposed method consists of the following two modules. First, it predicts what the topological information will be when the agent changes the viewing direction. Second, it predicts whether the predicted topological information will be highly rewarded in the next step. Each module is described in detail below.

1) Prediction of Topological Information for the Next Step: The network f is trained to predict what objects will be in view when an agent turns from the current state to an angle θ and what the resulting topological information will be. For example, as shown in Fig. 5, the green agent can see the black landmark even though the red agent cannot. Based on the information from the green agent, the red agent can predict that the black landmark will be visible if it faces a certain angle θ. The state information for the last $T = 3$ steps is input to the network f. Let $\mathbf{p}_i^t$, $\mathbf{d}_i^t$, and V_i^t denote the position, looking direction, and topological information of each agent at step t, respectively. And, let S_i^t be self topological information of a_i at step t. The network f takes as input the combination of these and the angle information θ for the next direction and outputs the predicted $\hat{S}_i^{t+1}$ of the self topological information at the next step $t+1$ for that agent.

$$\hat{S}_i^{t+1} = f\left(\{\{\mathbf{p}_i^{t'}\}_{i\in a_c}, \{\mathbf{d}_i^{t'}\}_{i\in a_c}, V_i^{t'}\}_{t'=t-T+1}^{t}, \theta\right). \quad (2)$$

Here, a_c refers to the group of agents, including itself, with whom a_i can communicate, and the position and visual direction of a_c are also utilized as inputs. $\hat{V}_i^{t+1}$ is then computed by substituting $\hat{S}_i^{t+1}$ for S_i^t in V_i^t. The network f learns to minimize the following loss (Mean Square Error).

$$\mathcal{L}_f = \frac{1}{N}\sum_{i=1}^{N} \|S_i^{t+1} - \hat{S}_i^{t+1}\|^2. \quad (3)$$

Topological information is difficult to compute for landmarks or adversarial agents whose positions are unknown. Particularly in multi-agent environments and in OWC where agents do not have constant communication with all other agents, constructing rules becomes increasingly complex as the number of agents increases. Therefore, it would be useful to train these in a neural network.

2) Prediction of Whether Topological Information Will Lead to Higher Rewards: Let r_i^t be the reward of step t for agent a_i, and define the network g that predicts the likelihood that the predicted topological information will lead to a higher reward than the current one by the following formula.

$$P(r_i^{t+1} - r_i^t > \Delta r) = g\left(\{\{\mathbf{p}_i^{t'}\}_{i\in a_c}, \{\mathbf{d}_i^{t'}\}_{i\in a_c}, \hat{V}_i^{t'}\}_{t'=t-T+1}^{t}\right). \quad (4)$$

The parameter Δr determines how much higher reward to expect from the current reward. If no other landmarks are visible to the agent or the same landmark is seen by the agent, as shown in Fig. 5 (right), the agent is unlikely to get a higher reward in the next action step. On the other hand, when different landmarks are visible, the probability of obtaining a high reward in the next action step is high. It is similar to learning the appropriate stance or formation for the task by tying it to topological information.

By learning these networks separately, the agent can predict the next topological information using f for each angle, and then use g to predict whether that topological information will yield a higher reward in the next step. Then, agent selects the most rewarding visual direction using end-to-end network. Letting $q_i = P(r_i^{t+1} > r_i^t)$, the network g is trained to minimize the following loss (Binary Cross Entropy).

$$\mathcal{L}_g = -\frac{1}{N}\sum_{i=1}^{N} y_i \log q_i + (1-y_i)\log(1-q_i), \quad (5)$$

where $y_i = 1$ if $r_i^{t+1} - r_i^t > \Delta r$ and $y_i = 0$ otherwise.

D. Training Algorithm

The training algorithm for this method is shown in Algorithm 1. During training, the agent learns what kind of observation and action policy π is rewarded by learning the usual Actor-Critic training. Each agent stacks the observed 1D RGB information of the past three frames and extracts it into a feature vector using a 1D convolutional network. If communication is possible, the communication signal is concatenated with the

Algorithm 1: Training.

for episode $= 1$ to N_{ep} **do**
  Receive initial state $\mathbf{x} \leftarrow \{\mathbf{p}_i, \mathbf{d}_i\}_{i=1}^{N}$
  for $t = 1$ to *max-episode-length* **do**
    for each a_i, select viewing direction θ
    Execute actions and get new state $\mathbf{x}' \leftarrow \{\mathbf{p}_i, \mathbf{d}_i\}_{i=1}^{N}$
    Get communication signals $\mathbf{x}' \leftarrow \{\mathbf{x}', \{V_i\}_{i=1}^{N}\}$
    for each a_i, select movement action using π
    Execute actions, get reward r, and get new state $\mathbf{x}'' \leftarrow \{\mathbf{p}_i, \mathbf{d}_i\}_{i=1}^{N}$
    Get communication signals $\mathbf{x}'' \leftarrow \{\mathbf{x}'', \{V_i\}_{i=1}^{N}\}$
    Store $(\mathbf{x}, \theta, a, r, \mathbf{x}', \mathbf{x}'')$ in replay buffer $\mathcal{D}$
    $\mathbf{x} \leftarrow \mathbf{x}''$
    for agent $i = 1$ to N **do**
      Sample a random minibatch of size S from $\mathcal{D}$
      Update critic (same as MADDPG [15])
      Update actor (same as MADDPG [15])
      Update f by minimizing the loss in (3)
      Update g by minimizing the loss in (5)
    end for
    Update target network parameters for each agent a_i
  end for
end for

Algorithm 2: Execution.

for episode $= 1$ to benchloopM **do**
  Receive initial state $\mathbf{x} \leftarrow \{\mathbf{p}_i, \mathbf{d}_i\}_{i=1}^{N}$
  for $t = 1$ to *max-episode-length* **do**
    for agent $i = 1$ to N **do**
      Get V_i
      for $j = 1$ to N_D **do**
        Set $\theta = 2\pi j / N_D$
        Estimate $\hat{S}_i^{t+1}$ by (2) and compute $\hat{V}_i^{t+1}$
        Calculate $P(r_i^{t+1} > r_i^t)$ by (4)
        RewardList$_i[j] \leftarrow P(r_i^{t+1} > r_i^t)$
      end for
      Select view dir. $j_{\max} = \text{argmax}_j(\text{RewardList}_i[j])$
    end for
    Change viewing direction to $2\pi j_{\max}/N_D$
    Select movement action for all agents
    Execute actions
  end for
end for

feature vector. The overall learning method uses the centralized training method same as in MADDPG [15]. The feature vector extracted from the Actor's own local observations is used to output the distance to move. The direction to be moved at this time is the direction after the visual direction change. The visual direction can be selected randomly or by the method described in Algorithm 2. Critic takes the combined feature vectors of all agents as input and outputs the Q-value through MLP during training.

At the same time as learning Actor-Critic, the prediction network f of topological information and the prediction network g of topological reward information described in Section III-C are learned from the data obtained here.

The execution algorithm is shown in Algorithm 2. At execution, the above two networks are used to predict the next topological information and the reward probability at that time from the θ candidates quantized every $2\pi/N_D$ when the visual direction is changed. By doing so, the change of visual direction is made in the direction with the highest probability of reward. In this letter, we set $N_D = 36$.

IV. Simulation Experiment

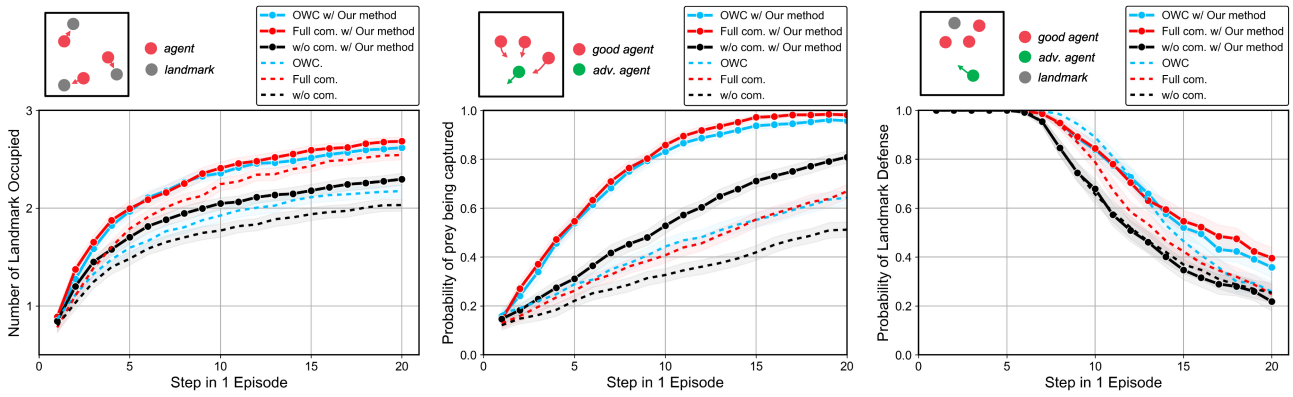
A. Settings

We investigated the effect of communication methods using OWC and visual coordination on performance in multi-agent reinforcement learning. For the multi-agent environment, we employed the Multi-agent Particle Environment (MPE) [15], which provides a variety of cooperative, adversarial, or communicative environments in the 2-D plane. Here, we conducted simulations to verify the performance of multi-agent reinforcement learning utilizing the visual coordination method described in Section III and optical wireless communication across multiple tasks. Note that in the case of visual coordination without communication setting, the method described in Fig. 5 is executed using only individual information without utilizing information from other agents. The case without visual coordination was defined as baseline, and in this case, the control action in the visual direction and the distance moved were combined as action. In other words, in this case, no prediction of topological information and rewards based on the next visual direction, as shown in Fig. 5, is performed. We finally confirmed whether the proposed method improves performance when applied to multiple algorithms in multi-agent reinforcement learning.

In addition to the *Simple Spread* task shown as an example in Section III-A, we performed benchmarking on the *Predator-Prey* task and *Keep-Away* task, which are provided as MPE. The *Predator-Prey* is a task where multiple good agents cooperatively chase and capture an escaping adversarial agent. The *Keep-Away* is a task where multiple good agents protect a landmark from an adversarial agent. For these task, topology information about the adversary agent is also added. On the other hand, communication is possible only among good agents. In this experiment, the processing part of visual information shared by both Actor and Critic is composed of four-layer 1D-convolutional with 64 filters and a kernel size of 4. The MLP layers of Actor and Critic are composed of a five-layer MLP with 256 units. In addition, the f and g networks are constructed with four-layer MLP with 256 units. In all experiments, we used the Adam optimizer with learning rates of 0.001 for both the Actor and Critic, and 0.0005 for the f and g networks. Additionally, Δr was set to 0.001 for all tasks.

B. Results

1) Simulation Results for Each Task: Fig. 6 shows the results for each task for a 200 K steps trained model, benchmarked 20 steps per episode for a total of 10 K times. First, the results



(a) *Simple Spread* (b) *Predator-Prey* (c) *Keep-Away*

Fig. 6. Benchmark results for each task per 1 episode (= 20 steps). The dashed line indicates the results of the baseline approach, while the solid line shows the results of the proposed method. The results of OWC using the proposed method are close to the results of Full Comm.

TABLE I
PERFORMANCE COMPARISON OF EACH LEARNING METHOD

	Full comm.		*OWC*		*w/o comm.*	
	Ours	Base.	Ours	Base.	Ours	Base.
MADDPG	90±0.79	85±0.86	87±0.75	72±1.1	81±0.92	67±1.1
DDPG	67±1.2	34±1.2	68±1.2	35±1.1	46±1.1	40±1.1
R-MADDPG	88±0.91	76±0.81	89±0.81	52±1.2	81±0.99	49±1.2
MARL-TRANS	72±8.6	53±3.8	71±7.5	26±3.5	58±7.1	37±6.2

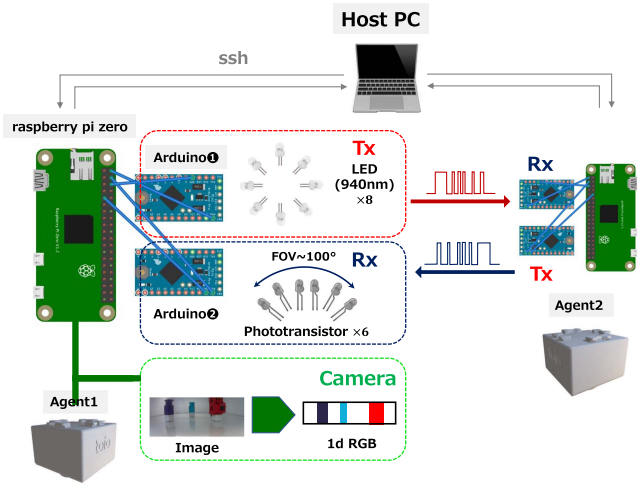


Fig. 7. Overview of the experimental setting with physical equipment.

TABLE II
ARRIVAL RATE COMPARISON ON PHYSICAL EQUIPMENT

Method	Arrival rate (20 steps)	Arrival rate (10 steps)
OWC	85%	73%
Full comm.	87%	83%
w/o comm.	66%	53%

of the *Simple Spread* task are shown in Fig. 6(a), where the performance is consistently higher with visual coordination than without visual coordination in all cases. This demonstrates the effectiveness of the proposed visual coordination method regardless of the communication method, which implies that the topological information shared by other agents in this method has an important influence on the accuracy of the visual direction control. Regarding the communication method, the performance is higher when OWC communication is used than in the case with no communication.

More interestingly, by including visual cooperative behavior learned in the proposed method, the performance of OWC becomes comparable to *Full comm.*

The results of the *Predator-Prey* task and the *Keep-Away* task are shown in Fig. 6(b) and (c). For these two tasks, we compared the results between training both good and adversarial agents using the proposed method and training only an adversarial agent using the proposed method. In other words, good agents are trained without the proposed method in the latter case. Fig. 6(b) demonstrates probability of adversarial agents captured by the good agents on the vertical axis, and (c) demonstrates probability of the good agent defending landmark from approaching adversarial agents during one episode on the vertical axis. First, in the *Predator-Prey* task, it is observed that the probability of reaching the adversarial agent is increasing regardless of the communication method using the visual coordination method. Furthermore, similar to the *Simple Spread* task, both *Full comm.* and *OWC* achieve comparable performance, capturing the adversarial agent with a probability of over 90% in a single episode. In the *Keep-Away* task, *Full comm.* and *OWC* also increase the probability of preventing the adversarial agent from reaching the landmark. In this way, the proposed method has been confirmed to improve performance in adversarial tasks, and to improve the effect of communication including OWC.

2) Simulation Results With Different RL Algorithms: In order to verify the improvement of the proposed method compared to *MADDPG* [15], we used *DDPG* [12], *R-MADDPG* [22], and *MARL-TRANS* [1] as baselines. The characteristics of each are shown below.

- *DDPG:* Unlike MADDPG, this uses decentralized training using only individual observations without using centralized training.

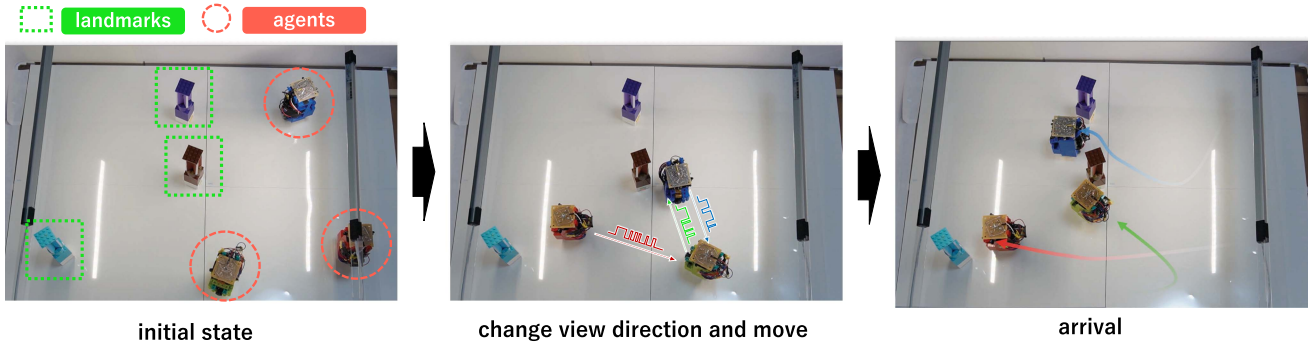


Fig. 8. Exemplar images of the physical devices operating with the proposed method with OWC. In the *Simple Spread* task, the agents share information using OWC (middle) and each agent decides its visual direction and moves towards the landmark (right).

- *R-MADDPG:* This approach introduces an LSTM layer into MADDPG to handle temporal information. The LSTM layer is introduced into the output of the convolutional layers of both the Actor and the Critic.
- *MARL-TRANS:* This approach introduces an attention mechanism into the communication content, under the assumption that the communication range is limited in distance. Note that Proximal Policy Optimization (PPO) [19] is used for the learning algorithm.

Table I shows the results when each algorithm is applied to the proposed method in the *Simple Spread* task. In the setting of $N = 3$ agents, DDPG's performance tends to degrade without centralized training. The proposed method improves its performance, while OWC achieves the same level of performance as *Full Comm*. Also, R-MADDPG is worse than MADDPG. Unlike the original letter where the state information such as positional data is directly inputted to the LSTM layer as observation, we introduced the LSTM layer after the CNN layer. Therefore, the information such as hidden state becomes stale as the CNN layer is updated [10]. Although detailed parameter tuning and optimization of the learning method were not performed in this study, the proposed method enhances the performance up to the same level as the original letter. Furthermore, for MARL-TRANS, merely changing the communication method to OWC significantly decreases performance. However, applying the proposed method improves its performance to the same level as *Full Comm*. These results demonstrate that the proposed method can be applied to various algorithms.

V. Experiments on Physical Equipment

A. Settings

For a physical agent, we used a small device named toio[1]. These devices can acquire their location information by reading the dedicated sheet where patterns of location information are encoded using an optical sensor. An external camera was connected using a Raspberry Pi Zero[2], and an Arduino[3] was connected for the purpose of OWC. The field size is 90 cm wide by 60 cm deep. The physical agents we created are shown in Fig. 1 and the overall configuration is shown in Fig. 7. Details of each part are described in the following.

1) OWC Part: For OWC, LEDs in the 940 nm band are used for transmitter. Manchester coding was used for the communication signal, and a phototransistor sensitive to the 940 nm band was used for the receiver (L-51ROPT1D1). The receiver was designed to acquire a range of 100°, which is the same as the camera angle of view. The transmitter was connected to 8 LEDs so that it could transmit throughout the perimeter so that it could be seen from any direction. Note that during execution, the timing of communication is separated to avoid confusion between the communication signals.

2) Camera Part: A camera with a FOV of about 100° was used. We extract only one line of the acquired image as used in the simulation. In order to use the behavior policy obtained in simulation in the physical device, images are converted to the data that should be obtained in simulation by a Sim2Real fashion. We employ a four-layer one-dimensional convolutional neural network for Sim2Real transformation and train it using 3,000 data samples.

During the experiment, the raspberry pi zero mounted on toio is connected to the host PC via ssh and communicates with it to determine what actions to have it take. The actions are determined based on the local observation information of each agent.

B. Results

Physical experiments were conducted with *Full comm.*, *OWC*, and *w/o comm.* First, the performance of OWC itself was checked. We confirmed the success of OWC communication within a range of approximately 80 cm in distance and a field of view of around 100 degrees by varying the relative angle and distance from the transmitter to the receiver's center.

Next, we confirmed the applicability of OWC to multi-agent operations. In this experiment, the communication signal for determining the visual direction is sent and received by OWC, and the other basic operation procedures are the same as in the simulation. As mentioned above, when using OWC, a tuning period is required to search whether the phototransistor can

[1]https://toio.io/
[2]https://www.raspberrypi.org/
[3]https://www.arduino.cc/

acquire the signal from the LED or not. During the tuning period, the Tx side rotates by 20° at regular intervals and the Rx side prepares to receive. Therefore, signal information is provided during the multi-agent operation if the signal is acquired as a result of tuning to shorten the operation time. In the experiments, benchmarking was performed based on the ratio of the number of landmarks reached in the final step per episode. The number of experiments was 10 episodes each, and the average percentage of the landmarks reached was calculated in the final (20th) step and the middle (10th) step. The results are summarized in Table II. The results reproduced that the operation with full communication achieved the highest performance and fastest arrival, while the performance decrement with OWC was relatively small. The exemplar operation is shown in Fig. 8.

VI. Conclusion

We propose a method for visual coordination in a multi-agent system with limited vision using OWC and evaluated its performance through simulations and physical experiments. By sharing information about what each agent is looking at, our method enables agents to imagine areas they cannot see and predict the probability of obtaining high rewards. Simulation results demonstrated that the OWC method outperformed the no-communication method and that controlling the viewing direction improved performance in all cases. Furthermore, combining OWC with the control of viewing direction achieved performance comparable to those of full communication. These results were also confirmed on a physical device. However, the current system has limitations, such as the need for optical axis alignment due to the use of only a few photodetectors. We anticipate that this challenge can be addressed in the future by using high-speed image sensors, such as photodiode arrays or event-based vision cameras, in the light-receiving portion of OWC.

References

[1] A. Agarwal, "Learning transferable cooperative behavior in multi-agent teams," Master's thesis, Carnegie Mellon University, Pittsburgh, PA, USA, May 2019.

[2] B. Baker et al., "Emergent tool use from multi-agent autocurricula," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. P1–P28.

[3] G. Chen et al., "A novel visible light positioning system with event-based neuromorphic vision sensor," *IEEE Sensors J.*, vol. 20, no. 17, pp. 10211–10219, Sep. 2020.

[4] T. Delbrück, B. Linares-Barranco, E. Culurciello, and C. Posch, "Activity-driven, event-based vision sensors," in *Proc. IEEE Int. Symp. Circuits Syst.*, 2010, pp. 2426–2429.

[5] Y. Du et al., "Learning correlated communication topology in multi-agent reinforcement learning," in *Proc. 20th Int. Conf. Auton. Agents MultiAgent Syst.*, 2021, pp. 456–464.

[6] J. N. Foerster, Y. M. Assael, N. de Freitas, and S. Whiteson, "Learning to communicate with deep multi-agent reinforcement learning," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 2145–2153.

[7] G. Gallego et al., "Event-based vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 154–180, Jan. 2022.

[8] R. Han, S. Chen, and Q. Hao, "Cooperative multi-robot navigation in dynamic environment with deep reinforcement learning," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 448–454.

[9] J. Jiang and Z. Lu, "Learning attentional communication for multi-agent cooperation," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 7265–7275.

[10] S. Kapturowski, G. Ostrovski, J. Quan, R. Munos, and W. Dabney, "Recurrent experience replay in distributed reinforcement learning," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. P1–P19.

[11] H. Komatsu et al., "The pointing performance of the optical communication terminal, SOLISS in the experimentation of bidirectional laser communication with an optical ground station," *Proc. SPIE*, vol. 11678, pp. 69–82, 2021.

[12] T. P. Lillicrap et al., "Continuous control with deep reinforcement learning," 2015, *arXiv:1509.02971*.

[13] Y.-C. Liu, J. Tian, N. Glaser, and Z. Kira, "When2com: Multi-agent perception via communication graph grouping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4105–4114.

[14] Y.-C. Liu, J. Tian, C.-Y. Ma, N. Glaser, C.-W. Kuo, and Z. Kira, "Who2com: Collaborative perception via learnable handshake communication," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 6876–6883.

[15] R. Lowe, Y. Wu, A. Tamar, J. Harb, O. P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6379–6390.

[16] L. Meng, T. Hirayama, and S. Oyanagi, "Underwater-drone with panoramic camera for automatic fish recognition based on deep learning," *IEEE Access*, vol. 6, pp. 17880–17886, 2018.

[17] I. Mordatch and P. Abbeel, "Emergence of grounded compositional language in multi-agent populations," *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, pp. P1495–P1502, 2018.

[18] N. Saeed, A. Celik, Y. T. Al-Naffouri, and M.-S. Alouini, "Underwater optical wireless communications, networking, and localization: A survey," *Ad Hoc Netw.*, vol. 94, 2019, Art. no. 101935.

[19] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.

[20] I. Takai, T. Harada, M. Andoh, K. Yasutomi, K. Kagawa, and S. Kawahito, "Optical vehicle-to-vehicle communication system using led transmitter and camera receiver," *IEEE Photon. J.*, vol. 6, no. 5, pp. 1–14, Oct. 2014.

[21] N. Ukita and T. Matsuyama, "Real-time cooperative multi-target tracking by communicating active vision agents," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2002, pp. 14–19.

[22] Rose E. Wang, M. Everett, and Jonathan P. How, "R-MADDPG for partially observable environments and limited communication," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. P1–P9.

[23] Z. Wang, Y. Ng, J. Henderson, and R. Mahony, "Smart visual beacons with asynchronous optical communications using event cameras," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2022, pp. 3793–3799.

[24] J. Wu, X. Sun, A. Zeng, S. Song, S. Rusinkiewicz, and T. Funkhouser, "Spatial intention maps for multi-agent mobile manipulation," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 8749–8756.

[25] T. Yamazato et al., "Image-sensor-based visible light communication for automotive applications," *IEEE Commun. Mag.*, vol. 52, no. 7, pp. 88–97, Jul. 2014.