A Cooperative Behavior Learning Control of Multi-Robot using Trace Information

Tomofumi Ohshita*, Ji-Sun Shin*, Michio Miyazaki** and Hee-Hyol Lee *
*Waseda University, **Kanto Gakuin University

Abstract: The distributed autonomous robotic system has superiority of robustness and adaptability to dynamical environment, however, the system requires the cooperative behavior mutually for optimality of the system. The acquisition of action by reinforcement learning is known as one of the approaches when the multi-robot works with cooperation mutually for a complex task. This paper deals with the transporting problem of the multi-robot using Q-learning algorithm in the reinforcement learning. When a robot carries luggage, we regard it as that the robot leaves a trace to the own migrational path, which trace has feature of volatility, and then, the other robot can use the trace information to help the robot, which carries luggage. To solve these problems on multi-agent reinforcement learning, the learning control method using stress antibody allotment reward is used. Moreover, we propose the trace information of the robot to urge cooperative behavior of the multi-robot to carry luggage to a destination in this paper. The effectiveness of the proposed method is shown by simulation.

Keywords: multi-agent systems, cooperative behavior, reinforcement learning, stress antibody allotment reward.

I. Introduction

The distributed autonomous robotic system has superiority of robustness and adaptability to dynamical environment, however, the system requires cooperative behavior mutually for optimality of the system. The acquisition of action by reinforcement learning is known as one of the approaches when the multi-robot works with cooperation mutually for a complex task.

To establish cooperative behavior, we should consider firstly, the uncertainty of state transition problem because of existence of more than one agent, secondly, the perceptual aliasing problem because of limitation for sensory input such as view, and thirdly, the reward sharing problem which is accurately distributing reward to indirect contribution for cooperation. In cooperative behavior of autonomous robots, Hong et al [1] put forward a cooperative behavior learning control using a stress antibody allotment reward in which the robots obtain a stress antibody to promote cooperative behavior of multi-robot.

This paper deals with transporting problem in multi-agent systems using Q-learning algorithm, and aims to establish of cooperative behavior for performing effective work. When a robot carries luggage, we regard it as that the robot leaves a trace to the own migrational path, which trace has feature of volatility, and then, the other robot can use the trace information to help the robot under carrying luggage. To solve problems mentioned above on multi-agent reinforcement learning, we use the learning control method using a stress antibody allotment reward. Moreover, we propose the trace information of the robot to urge cooperative behavior of the multi-robot to carry luggage to a destination in this paper.

We verify the influence on easing the perceptual aliasing problem, and show the effectiveness of the proposed method by simulation.

II. Q-learning Algorithm

1. Q-learning

Q-learning algorithm in reinforcement learning, which is classified in algorithm of the environmental identification type, is used in this paper. Q-learning has been devised by Watkins [6], and estimates the Q value which represents effectiveness of actions through interaction by trial and error with environment. An optimal action is easily obtained from the optimal value function $Q(s, a)$. The process is as follows:

```
Initialize $Q(s, a)$ arbitrarily;
Repeat (for each episode);
  Repeat (for each step of episode);
      Observe a state observation $s_t$;
      Choose $a_t$ from $s_t$ using policy derived from Q
       value;
```

```
        Take action $a_t$, observe $r_t$, $s_{t+1}$;
        Q value is updated by the following update
          equation;
            $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_t + \gamma \max_{a \in A} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]$   (1)
          where
            $\alpha$ : Learning rate   $(0 < \alpha \leqq 1)$
            $\gamma$ : Discount rate   $(0 \leqq \gamma < 1)$
        $t \rightarrow t+1$
      until $s_t$ is terminal;
    until all episodes are finished;
```

III. Transporting Problem

1. Grid World

The autonomous robots carry large and small luggage to a destination in this transporting problem. There is large and small luggage in the grid world. When a robot carries small luggage, a robot moves at 1/3 speed of normal movement speed. On the other hand, when a robot carries large luggage, the robot moves at 1/5 speed. Additionally, a robot carrying large luggage can move at 1/3 speed of normal movement speed by cooperating with other robot. When a robot carries small luggage, a reward r_s is given to the robot. On the other hand, when a robot carries a large luggage, a reward r_l is given to the robot. When a robot carries a large or small luggage to the destination, a reward r_g is given to the robot. When a robot encounters an obstacle or other robots, which cannot cooperate, negative reward r_o and r_e are given to the robot respectively.

2. Transfer Robot

A robot can recognize a cell in which the robot exists in the grid world and learn actions by Q-learning. The removable course of a robot is four ways; up, down, right and left. A robot acts as follows:

(1) A robot searches luggage or goal repeating movement by Q-learning algorithm;

(2) A robot has two kinds of learning module: CoopQL and CarriQL. The robot learns the movement to luggage and/or the cooperative behavior by CoopQL, and learns a movement to the destination by CarriQL after the robot maintains luggage;

(3) More than one robot cannot exist in the same cell;

(4) When a robot, which carries large luggage by oneself, meets with other robot which has nothing, the two robots can carry large luggage cooperatively;

(5) When a robot meets with other robot except of the case (4), the robot choices other ways according to the action selection method, a negative reward r_e is given to the robot;

(6) When a robot encounters obstacles, the robot choices other ways by the action selection method and a negative reward r_o is given; and

(7) When a robot is surrounded on all side by other robots and/or the obstacles, the robot waits for one turn to be over on the same cell.

IV. Cooperative Behavior Learning Control

1. Stress Antibody Allotment Reward

The stress antibody allotment reward [1] is used as a method to establish a cooperation behavior. Here, a robot under carrying large luggage undergoes a stress, and at the same time, produces an antibody against the stress, and then hands over it as a reward to the other robot, which supports carrying luggage. The cooperative behavior with other robot is promoted in such away. The cooperation by the stress antibody allotment reward is illustrated in Fig.1.

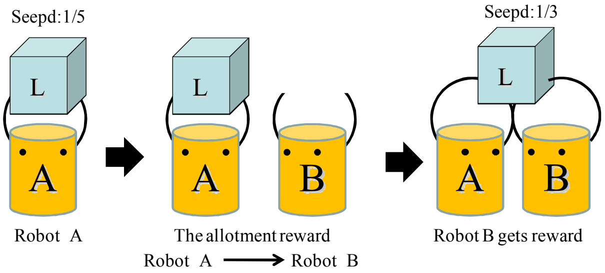


Fig.1 Cooperation by stress antibody allotment reward

2. Action-Value Function with Trace Information

When a robot carries luggage to the destination, the robot undergoes a stress and leaves volatile trace information (field sign) to own migration path. Additionally, the robot can recognize presence information on field sign in the cell, in which the robot exists. Thus, volatile trace information plays a role as rescue signal for the robot, which does not have luggage from the robot carrying large luggage. The volatilization rate is defined as follows:

$$fs(t)=\begin{cases}\sigma\ fs(t-1) & : \ (fs(t-1)\geq fs_0) \\ 0 & : \ (fs(t-1)<fs_0)\end{cases} \qquad (2)$$

where t is the number of the turns from putted on the grid world, σ is volatilization rate of the field sign ($0\leqq\sigma<1$). Thus, the action-value function includes an absolute coordinate of the robot existed, presence information on field sign on its cell, and actions of the robot. It is represented as follows.

$$Q(s_t, fs_{t,} a) \qquad (3)$$

$$s_t=(x_t, y_t), fs_t=\{0,1\}$$

V. Simulation

1. Situation Setting

A task of transporting large and small luggage to the destination is simulated in this paper. There are large and small 10 luggage, respectively, two walls blocking the way of the goal in the grid world, which size are 30 × 30 cells sizes, and a number of robots is four. One cell movement of a robot is called as one step, and one cell movement of all robots is called as one turn. The rewards given from environment are $r_g= 100$, $r_l= 60$, $r_s= 40$, $r_o= -5$, $r_e= -3$, $r_a= 30$. The initial value of field sign $fs(t_0)$ = 15, the volatilization rate of the field sign σ = 0.8, and the minimum value of filed sign fs_0 =1.0. Therefore, a robot can leave the field sign for up to three cells or less in the grid world. The ε-greedy rate is used as the action selection method. The ε-greedy rate of CoopQL and CarriQL are 0.25 and 0.75, respectively. The initial configuration is shown in Fig.2.

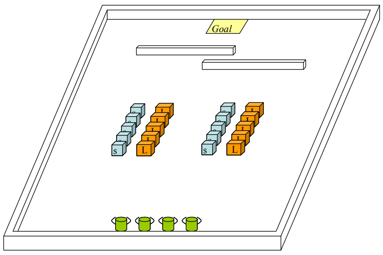


Fig.2 Initial configuration

2. Simulation

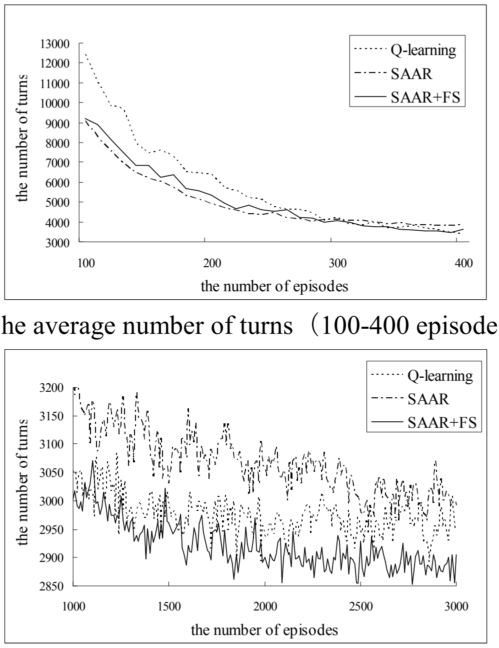
We attempt a comparison between proposed method using the stress antibody allotment reward and the field sign (SAAR+FS), the method using only stress antibody allotment reward (SAAR), and the method using only Q-learning (Q-learning) by simulation.

When robots finished carrying all luggage to the destination, it is called as one episode and one trial is 3000 episodes here. An average value of ten trials is used as the number of the turns. The average number of turns by each learning method is shown in Fig.3.

(1) Average number of turns of every ten episodes

As shown in Fig.3, a lot of average numbers of turns are needed at early period of learning in the Q-learning method. However, the average number of turns has decreased by one-quarter in SAAR method and SAAR+FS method. Moreover, SAAR+FS method has few number of turn in 1000-3000 episodes advanced by learning.

As the result, SAAR method is effective against reduction in the number of turns at early period of learning. Additionally, in SAAR method, there is an effect of decreasing the number of turns through entire learning process.



a) The average number of turns（100-400 episodes）

b) The average number of turns（1000-3000 episodes）

Fig.3 Comparison of the average number of turns by each learning method

(2) The total number of unnecessary encounters and cooperation

Fig.4 and Fig.5 show the total number of unnecessary encounters and cooperation in each learning method, respectively. The number of cooperation is ten times or less an episode, 30000 times or less 3000 episodes because large luggage is 10 pieces.

As shown in Fig.4, total number of unnecessary encounters for the SAAR method is much more than Q-learning method. This result is caused by increasing an unnecessary encounter due to crowd of robots by a reward given for the cooperative behavior. A robot can acquire a behavior to evade the obstacle, because a negative reward is given when the robot encounters obstacles. Thus, it seems many of unnecessary encounters with a robot at latter period of learning are occurred. On the other hand, the total number of unnecessary encounters in the SAAR+FS method is less than that in the SAAR method as well as the Q-learning method, regardless it uses the stress antibody allotment reward.

Furthermore, the total number of cooperation in SAAR+FS method is compared with the total number of cooperation in the SAAR method. It is shown that the total number of cooperation in the SAAR+FS method is more than that in SAAR method. On the other hand, Q-learning method is heavily less than that in the other two methods because a reward for the cooperation is not given.

As the results, it is shown that the robot can learn cooperative behavior to evade an unnecessary encounter by using the SAAR+FS method, because an unnecessary action is distinguished from a cooperative action. The SAAR+FS method can decrease the number of turns in a whole episode.

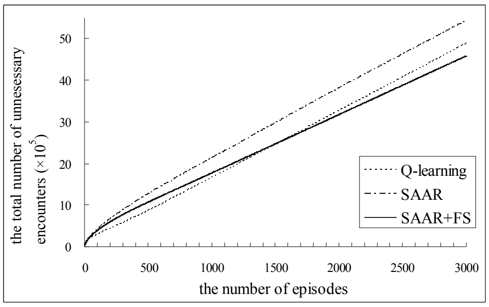


Fig.4 Comparison of the total number of unnecessary encounters by each learning method

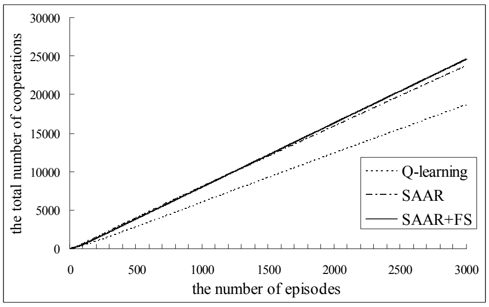


Fig.5 Comparison of the total number of cooperation by each learning method

VI. Conclusion

It is difficult to learn cooperative behavior accurately under environment with the uncertainty of state transition problem and the perceptual aliasing problem, above all, under circumstances in which a sensory input of a robot is limited only to the absolute coordinate aggravate the problems.

To relieve the problems, we used the stress antibody allotment reward, and proposed trace information of the robot to urge cooperative behavior. Effectiveness of the proposed method was verified by the transporting problem.

As the results of the simulation, the proposed method achieved a reduction in the number of turns of one-quarter at the level in early period of learning, and inhibited the number of turns in episodes advanced by learning. Furthermore, the proposed method can distinguish between rational actions and irrational actions, and then the useless step of needless encounter with other robots was decreased. The effectiveness of the proposed method was confirmed through the simulations.

References

[1] S.Hong, M.Miyazaki and H.Lee: "Learning Control of Carrier Robots with Cooperative Behavior", Electronics, Information and Systems Conference, IEEJ, MC4-1 (2006)

[2] S.Arai: "Multiagent Reinforcement Learning Frameworks: Steps toward Practical Use", Journal of JSAI, Vol.16, No.4, pp.476-481 (2001)

[3] S.Arai, K.Miyazaki and S.Kobayashi: "Methodology in Multi-Agent Reinforcement Learning: Approaches by Q-Learning and Profit Sharing", Journal of JSAI, Vol.14, No.4, pp.609-617 (1998)

[4] K.Miyazaki, S.Arai, S.Kobayashi: "Learning Deterministic Policies in Partially Observable Markov Decision Processes", Journal of JSAI, Vol.14, No.1, pp.148-156 (1999)

[5] R.S.Sutton and A.G.Barto: "Reinforcement Learning: An Introduction", The MIT Press (1998)

[6] C.J.C.H.Watkins and P.Dayan: "Technical Note: Q-Learnging", Machine Learning, Vol.8, pp.279-292(1992)