International Conference on Control, Automation and Systems 2007
Oct. 17-20, 2007 in COEX, Seoul, Korea

Cooperative Behavior Acquisition of Multiple Autonomous Mobile Robots by an Objective-based Reinforcement Learning System

Kunikazu Kobayashi[1], Koji Nakano[1], Takashi Kuremoto[1] and Masanao Obayashi[1]

[1]Division of Computer Science & Design Engineering, Yamaguchi University, Ube, Japan
(Tel : +81-836-85-9519; E-mail: koba@yamaguchi-u.ac.jp)

Abstract: The present paper proposes an objective-based reinforcement learning system for multiple autonomous mobile robots to acquire cooperative behavior. The proposed system employs profit sharing (PS) as a learning method. A major characteristic of the system is using two kinds of PS tables. One is to learn cooperative behavior using information on other agents' positions and the other is to learn how to control basic movements. Through computer simulation and real robot experiment using a garbage-collection problem, the performance of the proposed system is evaluated. As a result, it is verified that agents select the most available garbage for cooperative behavior using visual information in an unknown environment and move to the target avoiding obstacles.

Keywords: objective-based reinforcement learning, multiagent system, cooperative behavior, autonomous mobile robot, profit sharing

1. INTRODUCTION

Reinforcement learning is a method that agents will acquire the optimum behavior by trial and error by being given rewards in an environment as a compensation for its behavior [1], [2]. Most of studies on reinforcement learning have been done for a single agent learning in a static environment. The Q-learning which is a typical learning method is proved that it converges to an optimum solution for Markov Decision Process (MDP) [3]. However, in a multiagent environment, as plural agents' behavior may affect the state transition, the environment is generally considered as non Markov Decision Process (non-MDP), and we must face critical problems whether it is possible to solve [4].

On the above problems in a multiagent environment, Arai et al. have compared Q-learning with profit sharing (PS) [5] using the pursuit problem in a grid environment [6]. As a result, Q-learning has an instability for learning because it uses Q values of the transited state in an updating equation. However, PS can absorb the uncertainty of the state transition because of cumulative discounted reward. Therefore, they concluded that PS is more suitable than Q-learning in the multiagent environment [6], [7]. Uchibe et al. have presented the capability of learning in a multiagent environment since relation between actions of a learner and the others is estimated as a local prediction model [8]. However, PS has a problem of inadequate convergence because PS reinforces all the pairs of a state and an action irrespective of the achievement of a purpose [9].

The present paper presents an objective-based reinforcement learning system for multiple autonomous mobile robots to solve the above problem and to create cooperative behavior. The performance of the proposed system is verified through computer simulation and real robot experiment.

2. PROPOSED SYSTEM

2.1 Architecture

The present paper presents an objective-based reinforcement learning system as illustrated in Fig. 1. The proposed system is composed of three parts; an action controller, a learning controller and an evaluator. The feature of the system is to divide behavior of an agent into cooperative and basic behavior to learn separately. The learning of cooperative behavior is using information of the other agents' positions and the present state. The learning of basic behavior is to learn how to control own basic behavior like *go forward* or *turn right*. In a general learning method, when an agent acquires a reward it can hardly estimates own action whether it can cooperate or not. To solve this problem, the proposed system divides behavior into two kinds of behavior and each one is evaluated using different criteria.

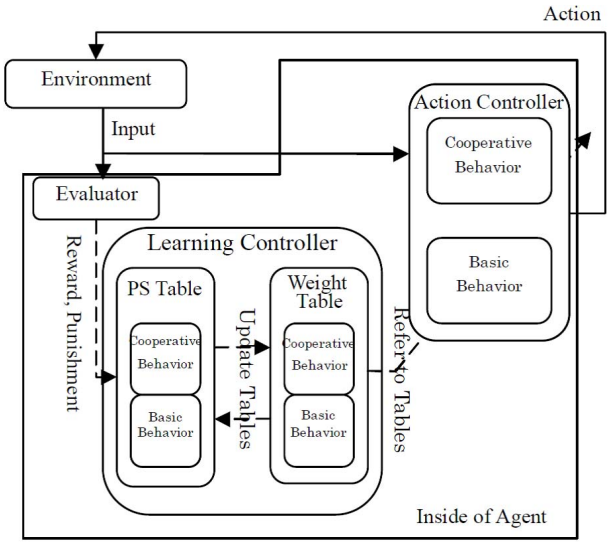


Fig. 1 Architecture of the proposed system.

978-89-950038-6-2-98560/07/$15

2.2 Action controller

An action is selected by the Boltzmann distribution. It is defined by the weight $w(s,a)$ of rules created by the pairs of a state s and an action a.

$$B(a|s) = \frac{e^{w(s,a)/T}}{\sum_{b \in A} e^{w(s,b)/T}}, \quad (1)$$

where $B(a|s)$ is a probability selecting action a at state s, T is a positive temperature constant and A is a set of available actions.

2.3 Learning controller

The PS is employed as a learning method for an agent.

$$w(s,a) = w(s,a) + f(t,r), \quad (2)$$

where t is a time, r is a reward and $f(\cdot,\cdot)$ is a reinforcement function. In the present paper, the following function is used as function f.

$$f(t,r) = r\gamma^{t_G - t}, \quad (3)$$

where γ is a decay rate and t_G is a time in the goal state. Equation (3) satisfies the rationality theorem of PS which guarantees successful convergence [10].

2.4 Evaluator

The different criteria are prepared for cooperative and basic behavior. This is because one can judge whether success and failure of agent's behavior come from cooperative behavior or basic behavior.

3. EXPERIMENT

The proposed system was applied to a garbage-collection problem which is one of the standard multi-agent tasks. The performance of the proposed system is evaluated through computer simulation and real robot experiment.

3.1 Problem setting

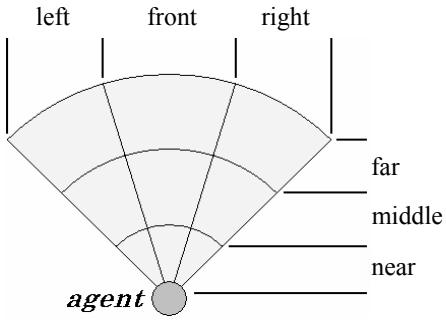
In a field, there are two agents, some garbage and one trash can, and then agents must collect all the garbage and take it to the trash can.

As shown in Fig.2, an input for the agent is classified into nine sub-states, combinations of three sorts of orientations (left, front or right) and three sorts of distances (near, middle or far).

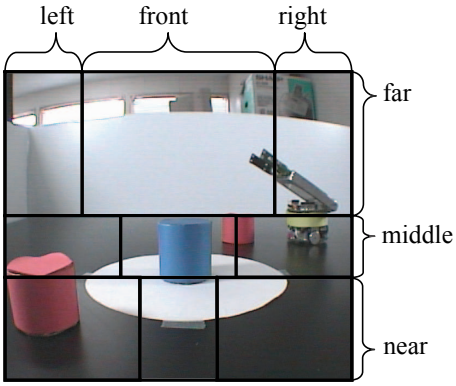
The action of agents is evaluated using four kinds of criterion as shown in Table 1. In this table, ◯ means reward or punishment is considered and × is not considered.

3.2 Computer simulation

A simulation field is a 21x21 grid world and there are 10 garbage, 2 agents and 1 trash can in the field as shown in Fig.3. One trial is defined as until all garbage are collected, and 100 trials are considered as one episode. The number of average steps is calculated after repeating 100 episodes. At this time, $w(s,a)$ are initialized for each episodes. To verify the effectiveness of the proposed system, it is compared with the standard PS system (conventional system).



(a) State classification in computer simulation



(b) State classification in real experiment

Fig. 2 State classification.

Table 1 Definition of reward and punishment.

condition	cooperative action	basic action
Reward: an agent arrives at the target garbage or the trash can	◯	◯
Punishment: an agent decide the same garbage with other agents	◯	×
Punishment: an agent bumps obstacles	×	◯
Punishment: an agent loses the target	×	◯

Figure 4 and Table 2 show the result of the computer simulation. In the case that one agent can observe the other, the agent using the proposed system learns faster than the agent using the conventional system. From this result, it is shown that the proposed system realizes cooperative behavior. However, when the agent is compared with the agent which is using the conventional system and do not observe the other agent, the performance of the proposed agent is similar to that of the conventional agent.

Figure 5 illustrates cooperative behavior observed in the experiment in which agent observes the other one. After agent 1 took garbage to the trash can (Fig.5(a)), it

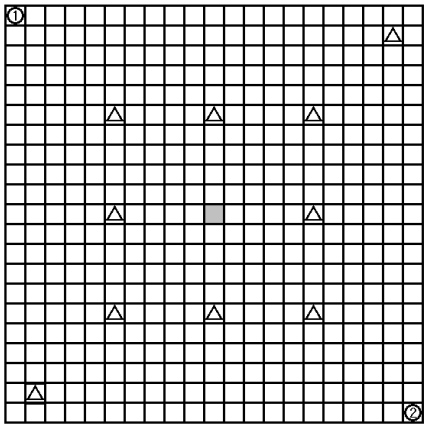


Fig. 3 Initial position of two agents (◯), ten garbages (△) and one trash can (shaded □).

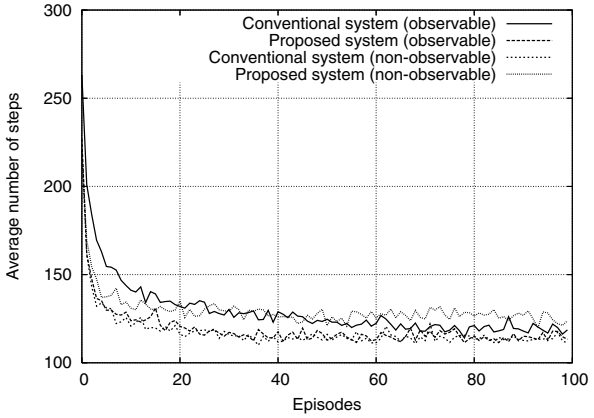
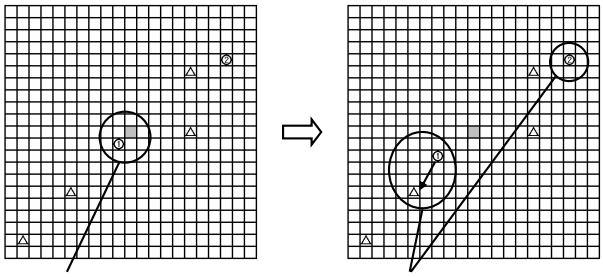


Fig. 4 Performance comparison of the proposed and conventional systems.

Table 2 The average number of steps in the final trial.

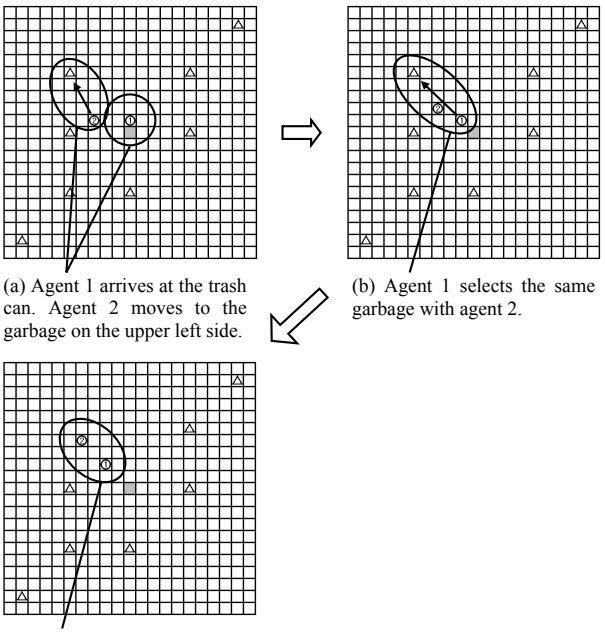
	observable	non-observable
conventional method	118.7	111.1
proposed method	113.3	123.8



(a) Agent 1 arrives at the trash can.

(b) Agent 1 moves to the garbage opposite to agent 2.

Fig. 5 An example of cooperative behavior acquired in the proposed system with observing the other agent.



(a) Agent 1 arrives at the trash can. Agent 2 moves to the garbage on the upper left side.

(b) Agent 1 selects the same garbage with agent 2.

(c) Agent 2 picks up the target garbage. Agent 1 increases the number of steps because of a waste of time-steps.

Fig. 6 An example of cooperative behavior acquired in the proposed system without observing the other agent.

do not select the garbage near agent 2 as the object, but another one opposite to agent 2 (Fig.5(b)). Such behavior often occurred after learning with observing the other agents. On the other hand, Fig.6 depicts cooperative behavior without observing the other agent. After agent 1 reached the trash can (Fig.6(a)), as it selected the garbage which are also targeted by agent 2 (Fig.6(b)), it is clear that the number of steps is increased because of a waste of time-steps (Fig.6(c)).

3.3 Real robot experiment

Two Khepera robots (K-Team) [11], an image processing board (IP7000BD), a color CCD camera (DCC-2010N) and a robot control PC are used in the experiment. An experiment field is a 1[m]×1[m] square surrounded by white walls and there are five garbage, two robots and one trash can as shown in Fig.7.

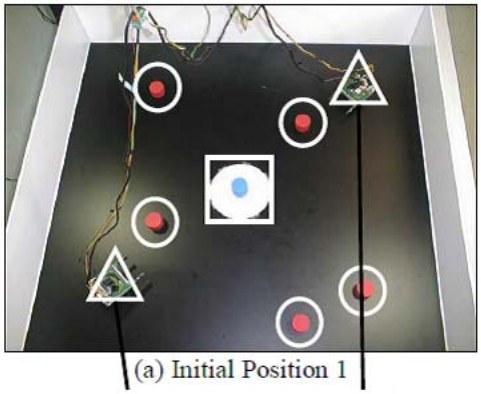
The following three kinds of the experiments were conducted to evaluate the learning ability of the proposed system.

Exp. 1: The robots are controlled using the learned weights in the simulation, which are not updated during Exp. 1.

Exp. 2: The robots are controlled using the learned weights in the simulation, which are updated during Exp. 2.

Exp. 3: The robots are controlled using the learned weights in Exp. 2 after the initial position of robots is changed.

Table 3 shows that the number of average steps in Exp. 2 is decreased compared with that in Exp. 1. Thus, the



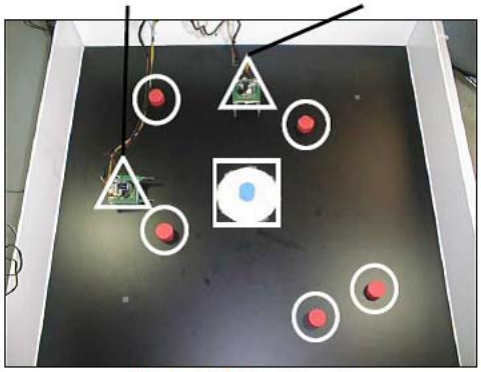


Fig. 7 Initial position of two agents (◯), five garbages (△) and one trash can (□).

learned weights in the simulation are available for the real robot environment, and furthermore, the proposed system can learn flexibly in real environment. On the other hand, the number of average steps in Exp. 3 is not increased compared with Exp. 2. Therefore, the weights learned in real environment are applicable to different environments, and this shows that the proposed system is robust.

Table 3 The average number of steps in Exp. 1 to 3.

	average number of steps
Exp. 1	201.9
Exp. 2	178.2
Exp. 3	161.1

4. SUMMARY

The present paper has proposed the objective-based reinforcement learning system for multiple autonomous mobile robots to acquire cooperative behavior. In the proposed system, robots select the most available target garbage for cooperative behavior using visual information in an unknown environment, and move to the target avoiding obstacles. The proposed system employs profit sharing (PS) and a characteristic of the system is using two kinds of PS tables. One is to learn cooperative behavior using information on other robot's positions, the other is to learn how to control basic movements. Through computer simulation and real robot experiment using a garbage-collection problem, it was verified that the proposed system is effective compared with the conventional system.

REFERENCES

[1] L. P. Kaelbling, M. L. Littman and A. P. Moore, "Reinforcement learning: A survey," *Journal of Artificial Intelligence Research*, Vol.4, pp.237-285, 1996.

[2] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, 1998.

[3] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Machine Learning*, Vol.8, pp.279-292, 1992.

[4] P. Stone and M. Veloso, "Multiagent systems: A survey from a machine learning perspective," *Autonomous Robots*, Vol.8, No.3, pp.345-383, 2000.

[5] J. J. Grefenstette, "Credit Assignment in Rule Discovery Systems Based on Genetic Algorithms," *Machine Learning*, Vol.3, pp.225-245, 1988.

[6] S. Arai, K. Miyazaki and S. Kobayashi, "Generating Cooperative Behavior by Multi-Agent Reinforcement Learning," *Proceedings of the 6th European Workshop on Learning Robots*, pp.143-157, 1997

[7] K. Miyazaki and S. Kobayashi, "Learning Deterministic Policies in Partially Observable Markov Decision Processes," *Proceedings of International Conference on Intelligent Autonomous System*, pp.250-257, 1998

[8] E. Uchibe, M. Asada and K. Hosoda, "State Space Construction for Cooperative Behavior Acquisition in the Environments Including Multiple Learning Robots," *Journal of the Robotics Society of Japan*, Vol.20, No.3, pp.281-289, 2002 *(in Japanese)*.

[9] K. Nakano, M. Obayashi, K. Kobayashi and T. Kuremoto, "Cooperative behavior acquisition for multiple autonomous mobile robots," *Proceedings of the Tenth International Symposium on Artificial Life and Robotics (AROB2005)*, CD–ROM, 2005.

[10] K. Miyazaki, M. Yamamura and S. Kobayashi, "On the Rationality of Profit Sharing in Reinforcement Learning," *Proceedings of the 3rd International Conference on Fuzzy Logic, Neural Nets and Soft Computing*, pp.285-288, 1994

[11] http:www.k-team.com