Cooperative Behavior Acquisition in Multi-agent Reinforcement Learning System Using Attention Degree

Kunikazu Kobayashi[1], Tadashi Kurano[2],
Takashi Kuremoto[2], and Masanao Obayashi[2]

[1] Aichi Prefectural University, 1522-3 Ibaragabasama, Nagakute, Aichi 480-1198, Japan
[2] Yamaguchi University, 2-16-1 Tokiwadai, Ube, Yamaguchi 755-8611, Japan
kobayashi@ist.aichi-pu.ac.jp,
{wu,m.obayas}@yamaguchi-u.ac.jp
http://www.ist.aichi-pu.ac.jp/~koba/

Abstract. In a multi-agent system, it becomes possible to solve a complicated problem by cooperative behavior with others. When people act in a group, as they are predicting the others' action, estimating the others' intention, and also making eye contact with others, they are realizing cooperative behavior efficiently. In the present paper, we try to introduce the concept of eye contact into a multi-agent system. In order to realize eye contact, we firstly define attention degrees both from self to the other and from the other to self. After that, we propose an action decision method that self agent makes easy to choose a target agent and to choose actions approaching to the agent using the attention degrees. Through computer simulation using a pursuit problem, we show that the agents making eye contact each other pursue preys by approaching each other. Simultaneously, we compare the proposed system with the standard Q-learning system and verify the usefulness of the proposed system.

Keywords: Multi-agent system, Cooperative behavior, Eye contact, Attention degree, Reinforcement learning, Pursuit problem.

1 Introduction

In multi-agent systems (MASs), intellectual behavior such as cooperative behavior can emerge toward a goal of agent group through mutual interaction among individual agents. In general, multi-agent systems have three major advantages over single-agent systems (SASs): robustness, flexibility, and load sharing [1]. As giving agents a reinforcement learning function, MASs can maximize its potential abilities such as problem solving and adaptation abilities [2,3].

To realize cooperative behavior in MASs, if agents are able to communicate with others using a highly accurate communication tool, agents can accurately obtain the other's action or intention [4]. Agents however has to predict the other's action or estimate the other's intention if agents are unable to communicate with others by restrictions of robot hardware and external environments.

T. Huang et al. (Eds.): ICONIP 2012, Part III, LNCS 7665, pp. 537–544, 2012.

Nagayuki et al. presented a policy estimation method which can estimate the other's action to be taken based on the observed information about the other's action sequence [5,6]. They successfully applied it to the reinforcement Q-learning method [7] and showed to get effective the other's policy. Meanwhile, Yokoyama et al. proposed an approach to model action decision based on the other's intention according to atypical situation such as human-machine interaction [8,9]. They presented three estimation levels of the other's intention and presented a computational model of action decision process to solve cooperative tasks through a psychological approach. In this context, Kobayashi et al. successfully presented an adaptive approach for automatically switching the above three estimation levels depending on the situation [10].

In the present paper, in order to realize cooperative behavior, we introduce a concept of eye contact, which is motivated by a method for detecting focusing intention of the learner in the collaborative learning environment [11]. Firstly, we formulate eye contact by a Q-value in reinforcement Q-learning method [7] and a P-value in the policy estimation method [5,6]. Secondly, we propose two kinds of attention degrees both from self to the other and from the other to self. Thirdly, we propose an action decision method that self agent makes easy to choose a target agent and to choose actions approaching to the agent using the attention degrees. Finally, through computer simulations using a pursuit problem, we show that the agents making eye contact each other pursue preys by approaching each other. Simultaneously, we compare the proposed system with the standard Q-learning system and verify the usefulness of the proposed system.

2 Reinforcement Learning

Reinforcement learning is a machine learning technique that a decision-making agent takes actions and then receives rewards in an environment, and finally acquires the optimum policy by trial and error [2,3].

The Q-learning method by Watkins et al. is a representative reinforcement learning technique and guarantees that a value function will converge to the optimal solution by appropriately adjusting a learning rate in Markov decision process environments [7]. A state-action value function $Q(s, a)$ (Q-value) is updated by (1) so as to take the optimal action by exploring it in a learning space.

$$Q(s, a) \leftarrow Q(s, a) + \alpha\delta, \tag{1}$$

where α is a learning rate ($0 < \alpha < 1$) and δ is a temporal difference error (TD error) denoted by (2).

$$\delta = r + \gamma \max_{b \in A} Q(s', b) - Q(s, a), \tag{2}$$

where r is a reward at the state s', s' is the next state after an agent takes action a, γ is a discount rate ($0 \leq \gamma \leq 1$), and A is a set of all possible actions.

Probabilistically, an agent selects action a at state s according to policy $\pi(s, a)$. Throughout the present paper, we employ the Boltzmann distribution defined by (3) as the policy.

$$\pi(s,a) = \frac{\exp\left(\beta Q(s,a)\right)}{\sum\limits_{b \in A} \exp\left(\beta Q(s,b)\right)}, \tag{3}$$

where β is a parameter to control randomness of action selection called as inverse temperature parameter. The policy $\pi(s,a)$ refers to a probability to select action a at state s.

3 Multi-Agent Systems

MASs have three major advantages over SASs: robustness, flexibility, and load sharing [1]. In higher dimensional space, however, MASs have so-called state-space explosion problem. The tile coding to be describe in Section 3.1 is well-known for solving the above problem. On the other hand, to realize cooperative behavior, we focus on predicting other's action to be described in Section 3.2 and attention degree to be proposed in Section 4.

3.1 Tile Coding

First of all, in the present paper, we consider a grid world as an external environment. To overcome state-space explosion problem in MASs, we firstly consider a generalization of state-space by random tile-coding [3]. As shown in Fig.1, the state-space is randomly covered by some square tiles in order to reduce the number of states.

By apply this random tile-coding to Q-learning, we get the following Q-value.

$$Q(s,a) = \sum_{i=1}^{n} q(i,a)\phi(i), \tag{4}$$

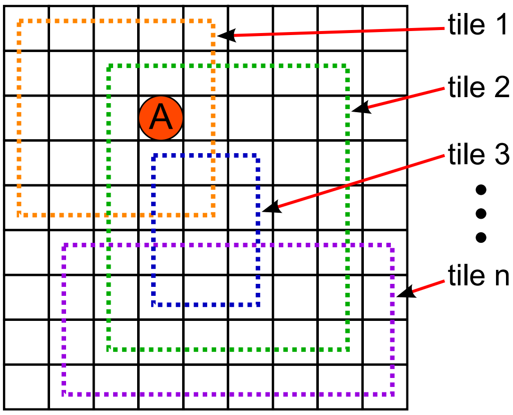


Fig. 1. Random tile coding (circle: agent, solid line: state space, dashed line: tiles)

where n denotes the number of tiles, $q(i, a)$ is the value function with respect to tile i and action a, and $\phi(i)$ is a binary function whether the agent exists in tile i or not as defined by:

$$\phi(i) = \begin{cases} 1 & \text{if the agent exists in tile } i, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

The q-value is updated by (6), which is similar with the update rule of Q-value in (1):

$$q(s, a) \leftarrow q(s, a) + \alpha\delta/n. \tag{6}$$

3.2 Policy Estimation

The policy estimation method can estimate the other's action to be taken based on the observed information about the other's action sequence [5,6]. The method predicts an other's action using a policy estimation function $P(s, a_o)$ (P-value). The P-value is updated by (7) for all the other's actions to be taken $a_o \in A$

$$P(s, a_o) \leftarrow (1 - \rho)P(s, a_o) + \begin{cases} \rho & \text{if } a_o = a_o^*, \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

where a_o^* is the actual other's action and ρ is a positive parameter ($0 \leq \rho \leq 1$). As updating P-value by (7), P-value with a_o^* increases and the other P-values decrease. Repeatedly updating P-values, an agent can predict other's actions. It should be noted that the following relation holds at any time:

$$\sum_{a_o \in A} P(s, a_o) = 1. \tag{8}$$

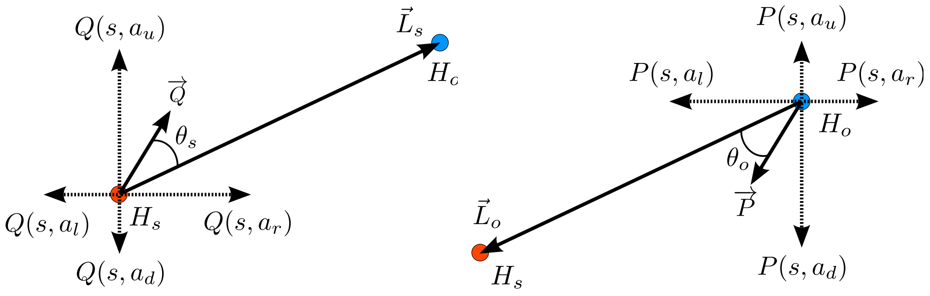
4 Intelligent Learning System Using Attention Degree

In order to emerge cooperative behavior, we introduce a concept of eye contact. It is motivated by a method for detecting focusing intention of the learner in the collaborative learning environment [11]. In the present paper, an intelligent learning system using attention degree is proposed. At first, we realize eye contact by attention degree (Section 4.1). Then, we propose an action decision method using attention degree (Section 4.2).

4.1 Attention Degree

First of all, we treat two kinds of attentions: attention from self to the other and attention from the other to self. Then, these two kinds of attentions are quantified by a Q-value in (1) and a P-value in (7). The attention degrees both from self to the other and from the other to self are illustrated in Fig.2. Note that both self agent H_s and the other agent H_o are placed on a grid world.

In Fig.2(a), we consider four Q-values at state s, i.e. $Q(s, a_u)$, $Q(s, a_d)$, $Q(s, a_l)$, and $Q(s, a_r)$. Here, a_u, a_d, a_l, and a_r refer to actions of moving up, down, left, and



(a) A resultant vector $\overrightarrow{Q}_s$ for H_s and a positional vector $\overrightarrow{L}_s$ for H_o.

(b) A resultant vector $\overrightarrow{P}_o$ for H_o and a positional vector $\overrightarrow{L}_o$ for H_s.

Fig. 2. Illustration of resultant and positional vectors

right, respectively. Then, we calculate a resultant vector $\overrightarrow{Q}$ and an angle θ_s between $\overrightarrow{Q}$ and a positional vector $\overrightarrow{L}_s$ for H_o. Finally, an attention degree from self to the other $AD(H_s, H_o)$ is defined by:

$$AD(H_s, H_o) = (\cos\theta_s + 1)/2. \tag{9}$$

It is clear that $AD(H_s, H_o)$ has value between 0 and 1, i.e. $AD(H_s, H_o) \in [0, 1]$. $AD(H_s, H_o)$ approaches to 1 as H_o pays attention to H_s.

Similarly, we consider four P-values at state s, i.e. $P(s, a_u)$, $P(s, a_d)$, $P(s, a_l)$, and $P(s, a_r)$ as shown in Fig.2(b). Then, we calculate a resultant vector $\overrightarrow{P}$ and an angle θ_o between $\overrightarrow{P}$ and a positional vector $\overrightarrow{L}_o$ for H_s. After that, an attention degree from self to the other $AD(H_o, H_s)$ is defined by:

$$AD(H_o, H_s) = (\cos\theta_o + 1)/2, \tag{10}$$

where $AD(H_o, H_s) \in [0, 1]$ holds.

4.2 Action Decision Using Attention Degree

To promote cooperative behavior, it is desired that the self agent approaches to the other agent having eye contact.

Using two attention degrees, i.e. $AD(H_s, H_o)$ in (9) and $AD(H_o, H_s)$ in (10), we choose a target agent t_a by

$$t_a = \arg\max_{i \in T_a} \frac{AD(H_i, H_s) \times AD(H_s, H_i)}{d(H_s, H_i)}, \tag{11}$$

where T_a refers to a set of subscript of other agents and $d(H_s, H_i)$ is a normalized distance between self H_s and other agent H_i ($i \in T_a$), i.e. $d(H_s, H_i) \in [0, 1]$.

In order to approach the self agent to the other agent having a higher attention value, Q-values should be recalculated by

$$Q'(s, a_k) = Q(s, a_k) \times (\cos\theta_k + 1)/2, \tag{12}$$

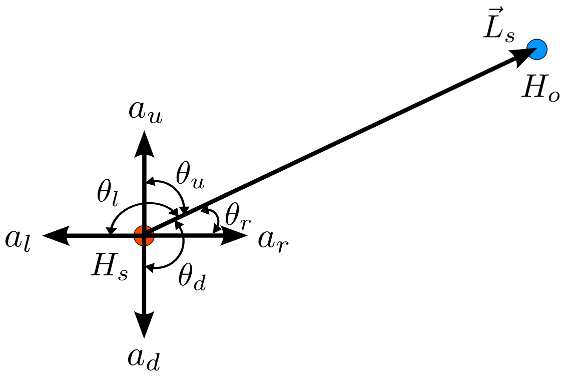


Fig. 3. Angeles between a positional vector $\overrightarrow{L}_s$ and directional vectors of action a_k

where a_k represents one of four possible actions in a grid world: a_u, a_d, a_l, or a_r and θ_k is an angle between a positional vector $\overrightarrow{L}_s$ and a directional vector of action a_k. θ_k is illustrated in Fig.3.

After that, agents select their actions based on the Boltzmann distribution (3).

Using the above action decision method, we expect to emerge cooperative behavior because agents try to cooperate with the other agent having eye contact.

5 Computer Simulation

In this section, through computer simulations using a pursuit problem, we verify the performance of the proposed intelligent learning system. At first, we describe a problem setting of the pursuit problem in Section 5.1. Secondly, we present a simulation setting in Section 5.2. Finally, we show simulation results in Section 5.3.

5.1 Problem Setting

A pursuit problem is a well-known multi-agent problem which plural hunters pursuit preys (or a prey) and catch them in a grid field. The followings are assumed in the present paper.

- Two dimensional 15×15 grid field with a torus structure.
- Six hunters H_i ($i \in \{1, 2, \cdots, 6\}$) and three preys P_j ($j \in \{1, 2, 3\}$) in the field. Initially, they are located randomly in the field.
- All the hunters can observe all the cells (complete observation) and act according to their own policy. The hunters can synchronously move up, down, left, or right by one cell, or stay on the same cell.
- All the preys can synchronously act according to a predefined policy.
- A goal state is assumed that each prey is occupied by any two hunters in two of four adjacent cells (up, down, left, and right). The two hunters can get a reward and the captured prey is removed from the field.

5.2 Simulation Setting

The hunters get a positive reward $r = 100$ if a goal state is reached. The number of steps is limited to 5,000 and we start a next trial if it reaches the limit.

The parameters were selected as learning rate $\alpha = 0.01$, discount rate $\gamma = 0.8$, the number of tiles $n = 1,500$, and positive parameter $\rho = 0.6$. The inverse temperature parameter was calculated by $\beta = 5.0 \times 10^{-4}e^{-t/100}$. Initial Q-values and P-values were set to 0.1 and 0.2, respectively. These parameters were selected so as to get the best performance through preliminary computer simulations.

5.3 Simulation Results

The learning curves is shown in Fig.4. In the figure, the horizontal axis represents the number of episodes and the vertical axis is the averaged number of steps. In the simulation, the number of steps is averaged for 5 trials. In this figure, red and green lines show learning curves for the proposed and the conventional systems, respectively. Here, the conventional system refers to a standard Q-learning system with random tile-coding. Note that a standard Q-learning system without random tile-coding could not tackle the given pursuit problem. As seen in Fig.4, the proposed method converges adequately but the conventional method does not converge at all. We conducted many computer simulations changing parameters and situations. It is verified that the proposed attention system works well in any cases. We observed that the proposed system promotes cooperative behavior.

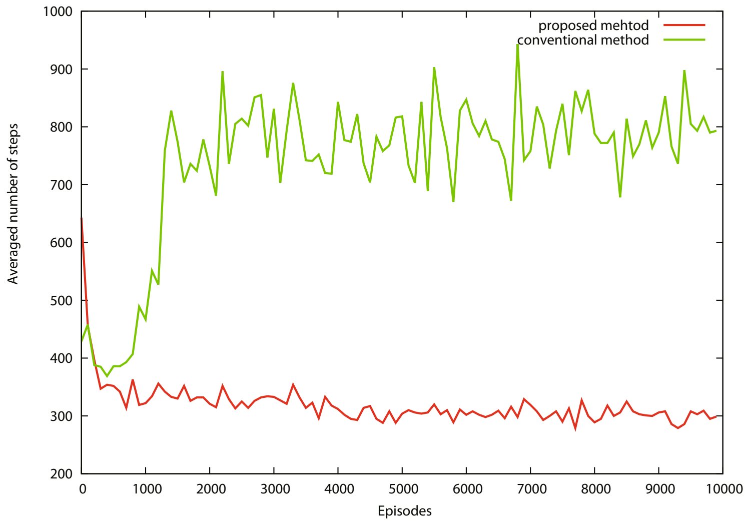


Fig. 4. Learning curves

6 Summary

In the present paper, we have proposed the proposed intelligent learning system using attention degree to emerge cooperative behavior in MASs. Firstly, we have introduced a concept of eye contact and formulated eye contact by a Q-value in reinforcement Q-learning method and a P-value in the policy estimation method. Secondly, we have proposed attention degrees both from self to the other and from the other to self. Thirdly, we have proposed an action decision method using the attention degrees that self agent approaches the other agent having eye contact. Finally, we have shown that the agents making eye contact each other pursue preys by approaching each other. Through computer simulations using a pursuit problem, we have verified that the proposed system has superior performance with the standard Q-learning system.

Acknowledgments. This work was partly supported by Grant-in-Aid for Scientific Research (No.23500181) from JSPS, Japan.

References

1. Stone, P., Veloso, M.: Multiagent Systems: A Survey from a Machine Learning Perspective. Autonomous Robots 8(3), 345–383 (2000)
2. Kaelbling, L.P., Littman, M.L., Moore, A.P.: Reinforcement Learning: A Survey. Journal of Artificial Intelligence Research 4, 237–285 (1996)
3. Sutton, R. S., Barto, A. G.: Reinforcement Learning: An Introduction. MIT Press (1998)
4. Bratman, M.E.: Intention, Plans and Practical Reason. Harvard University Press (1987)
5. Nagayuki, Y., Ishii, S., Ito, M., Shimohara, K., Doya, K.: A Multi-Agent Reinforcement Learning Method with the Estimation of the Other Agent's Actions. In: Proceedings of the Fifth International Symposium on Artificial Life and Robotics, vol. 1, pp. 255–259 (2000)
6. Nagayuki, Y., Ito, M.: Reinforcement Learning Method with the Inference of the Other Agent's Policy for 2-Player Stochastic Games. Transactions on the Institute of Electronics, Information and Communication Engineers J86-D-I(11), 821–829 (2003) (in Japanese)
7. Watkins, C.J.C.H., Dayan, P.: Q-learning. Machine Learning 8(3-4), 279–292 (1992)
8. Yokoyama, A., Omori, T., Ishikawa, S., Okada, H.: Modeling of Action Decision Process Based on Intention Estimation. In: Proceedings of Joint 4th International Conference on Soft Computing and Intelligent Systems and 9th International Symposium on Advanced Intelligent Systems, vol. TH-F3-1 (2008)
9. Yokoyama, A., Omori, T.: Model Based Analysis of Action Decision Process in Collaborative Task Based on Intention Estimation. Transactions on the Institute of Electronics, Information and Communication Engineers J92-A(11), 734–742 (2009) (in Japanese)
10. Kobayashi, K., Kanehira, R., Kuremoto, T., Obayashi, M.: An Action Selection Method Based on Estimation of Other's Intention in Time-Varying Multi-agent Environments. In: Lu, B.-L., Zhang, L., Kwok, J. (eds.) ICONIP 2011, Part III. LNCS, vol. 7064, pp. 76–85. Springer, Heidelberg (2011)
11. Hayashi, Y., Kojiri, T., Watanabe, T.: Focus Support Interface Based on Actions for Collaborative Learning. Neurocomputing 73(4-6), 669–675 (2010)