

Social Interaction of Cooperative Communication and Group Generation in Multi-Agent Reinforcement Learning Systems

Kun Zhang

Dept. of System Design Engineering
Graduate School of Engineering
University of Fukui
Fukui, Japan
Email: kzhang@ir.his.u-fukui.ac.jp

Yoichiro Maeda and Yasutake Takahashi

Dept. of Human and Artificial Intelligent Systems
Graduate School of Engineering
University of Fukui
Fukui, Japan
Email: (maeda, yasutake)@ir.his.u-fukui.ac.jp

Abstract—Recently, researches on multi-agent systems (MAS) which autonomous agents are able to learn cooperative behavior are actively performed. It is necessary for social agents to interact each other in order to have excellent cooperative performance in MAS. But it is difficult to give appropriate coordination at the right time, not to mention the generation of group behavior. We have aimed at the group behavior generation of social agents who have excellent autonomous learning ability like human through cooperative communication between agents to acquire cooperative behavior. Social agents are able to change the environment states to individual states and communicate their cooperative behavior by sending a special signal, which can not be understood originally but could be understood by other agents through reinforcement learning (RL) after some learning processes. Furthermore, if the agents could communicate their cooperative behavior successfully, their group identity will be strengthened by exchanging rewards among them. As the learning process, social agents can get the better cooperative ability than normal reinforcement learning method through group identity. Lastly, social agents not only adjust themselves to environment but also affect other agents to generate intelligent group behavior.

Keywords—Reinforcement Learning; Multi-Agent Systems; Social Interaction; Cooperative Communication; Group Generation

I. INTRODUCTION

In multi-agent systems (MAS), massive autonomous agents are required to solve complex problems by the coordination and cooperation of other agents flexibly[1]. On the other hand, reinforcement learning (RL) is an algorithm attempts to explore a suitable behavior to maximize a performance measurement based on interacting with an environment autonomously by repeating trial and error. RL adjusts behavior to a dynamic environmental without explicit teaching signals except rewards based on evaluation of current state and action. MAS is expected that it is able to learn the cooperative behavior and the group strategy by the application of the RL to autonomous agents[2], [3].

Generally, RL is basically developed for a single agent and it is used in MAS aspect, problems such that perceptual aliasing problem, concurrent learning problem and reward

distribution problem occurred[4]. In a cooperative social environment, it is effective to learn by sharing sensing information, episodes, learned policies or exchanging advice. However, one of the main problems concerned learning in a MAS environment is how to benefit from mutual interaction during the learning process[5]. It is possible to develop cooperation models to unify partial action policies to create new policies which can be optimal ones[6]. But for social agents, we believe that it is necessary to communicate the cooperative behavior and to give appropriate coordination at the right time[7]. We have aimed at the group behavior generation of social agents who have excellent autonomous learning ability like human through cooperative communication between agents to acquire cooperative behavior[8].

In this paper, social agents are able to change the environment states to individual states and communicate their states with other agents by sending a new undefined signal. For other agents, they can not understand the meaning of that signal and just take it as a new state of that agent during the learning process. After massive trial and error of RL algorithm, under the hint of this undefined signal, agents could get a better optimal policy than normal RL method. Therefore, although other agents do not know the meaning of the undefined signal, the overall identity of their behavior is to achieve the task cooperating with the signal sending agent. In other words, because both sides made similar coordinate decision, it is possible to communicate their cooperative behavior by sending a new undefined signal that can not be understood originally but could be gradually understood by other agents after some learning processes.

In addition, the agents who regularly made the similar decision have an identity of cooperation like the social identity in our human society. Based on social identity theory, these agents create an in-group or self-categorization because they have stronger organization than agents of normal RL method. Therefore, we propose a method to cause group generation by exchanging rewards as group identity among themselves. All agents not only receive rewards from the environment but

also exchange the reward among each others based on the undefined signal hints and their decisions. It is possible for exchanging rewards among agents to reinforce the coordinative behavior autonomously when the expected actions appeared. Agents are able to acquire excellent cooperative group behavior and strength their group identity by taking reward exchanging as the social interaction strategy like human in society.

II. SOCIAL INTERACTION OF COOPERATIVE COMMUNICATION AND GROUP GENERATION

In the normal multi-agent reinforcement learning algorithm, each individual agent can only repeat trial and error through the interaction with environment. In this paper, social agents can communicate with other agents by sending an undefined special signal that can not be understood originally but could be understood by other agents after some learning processes. If the agents could communicate their cooperative behavior successfully, their group identity will be strengthened by exchanging rewards among them. Based on the group identity, group generation will be possible. The general idea of proposed Multi-agent RL systems as shown in Fig.1.

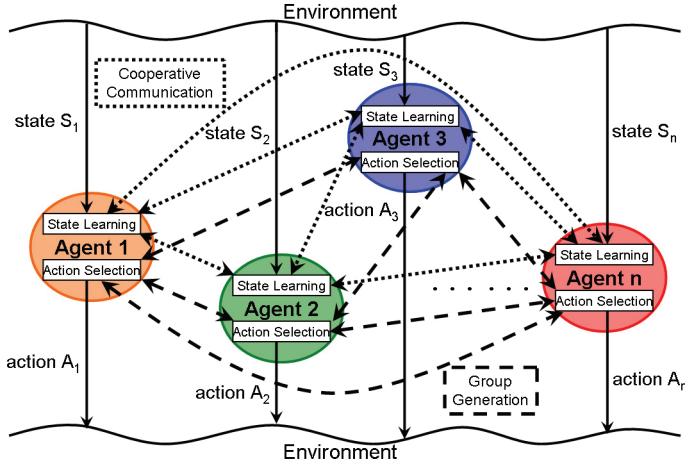


Fig. 1. Proposed Multi-agent RL systems

During the learning process, social agents can get the better cooperative ability than normal reinforcement learning method through the interaction of cooperative communication and group identity. Lastly, social agents not only adjust themselves to environment but also affect other agents to generate intelligent group behavior. Social agents will make decisions not only based on itself profit but also considering a joint profit with other agents.

A. Cooperative Communication Reinforcement Learning

For social agents, we believe that it is necessary to understand other behavior and to give appropriate coordination at the right time in MAS. In this cooperative communication reinforcement learning (CC-RL) algorithm, social agents are able to change the environment states to individual states and communicate their states with other agents by sending a new undefined signal. For other agents, they can not understand the

meaning of that signal and just take it as a new states during the learning process, and try all the possible state-action strategy by trial and error using RL algorithm.

In this approach, when the target was discovered, the agent who found will change the environment state of its perception range to the state of itself, and send a special signal information (we called Help) to other agents if the task could not be achieved by itself in a dynamic environment. But other agents are uncertain this Help meaning of this special signal information and record it as a new special state S_t^n not defined in their sates. In the algorithm, S_t^n is described by a flag (0 or 1) which means a task could be achieved or not by itself.

In addition, the Help signal emerge only when an agent found target and can not achieve it. So this S_t^n information is also related to target discovering or task performing. By this way, agents can perceive other agents' states and have to adjust this new state by learning. If the overall identity of their behavior is to achieve the task cooperating with the signal sending agent, the communication of their cooperative behavior becomes possible by sending a new undefined signal.

Concretely, once the target was discovered, the agent send Help information to other agents even they are outside the range of itself perception. At the same time, other agents receive the coordinates of generated Help and compute the relative direction from itself position to Help. The emergence of Help and the relative direction will be recorded as the environment changing information like normal RL method. All the agents will perceive the signal of Help through the communication among agents if one agent found target. Therefore, this CC-RL method based on the emergence of new signal and the relative direction can partly solve the problem of incomplete perception in MAS. In Q-Learning method, Q-value is estimated on the basis of experience as follows:

Cooperative Communication Reinforcement Learning

Initialize $Q_t(S_t^e \vee S_t^n, a_t)$ arbitrarily

Repeat (for each episode)

Initialize $S_t^e \vee S_t^n$

Repeat (for each step of episode)

Observe S_t^e

Cooperative communication of individual state

Receive S_t^n from other agents

Choose a_t from $S_t^e \vee S_t^n$ using policy derived from Q-value

Take action a_t

Observe r_t and next state $S_{t+1}^e \vee S_{t+1}^n$

$$Q_t(S_t^e \vee S_t^n, a_t) \leftarrow Q_t(S_t^e \vee S_t^n, a_t) + \alpha[r_t + \gamma \max_{a \in A} Q_t(S_{t+1}^e \vee S_{t+1}^n, a) - Q_t(S_t^e \vee S_t^n, a_t)]$$

$$S_t^e \vee S_t^n \leftarrow S_{t+1}^e \vee S_{t+1}^n$$

Until $S_t^e \vee S_t^n$ is terminal

where,

S_t^e : environment state of normal RL method,
 S_t^n : the new undefined signal information state informed by other agent in current moment t . It includes the relative direction from own position to this new signal position.

The action evaluation value $w(s, b)$ becomes like in Eq.(1) in Profit Sharing method.

$$w_t(s_t^e \vee s_t^n, b) \leftarrow w_t(s_t^e \vee s_t^n, b) + f(h) \quad (1)$$

$w_t(s_t^e \vee s_t^n, b)$ is the action evaluation value when the state is $s_t^e \vee s_t^n$ and the action is b in current moment t . Reinforcement function $f(h)$ is the same as one of normal Q-Learning method. Efficient cooperative behaviors will be expected under the communication of this new signal hint among agents.

B. Group Generation Reinforcement Learning

By sending a new signal to others in the approach of CC-RL, the communication among agents is possible. But it just tells the partial information to each other not giving any evaluation for the action under this information. In our human society, if person A likes an action performed by person B, then A will praise B. The possibility of perfect action of B will be repeated in next time for the praising force. Based on this idea, we believe that if social agents want to have excellent cooperative ability like human, the agent should be praised if its action accords with the expected behavior.

In this approach, at a certain time t if the agent A_1 discovered the target but it could not achieve the task with itself power, then A_1 sent a new signal to other agents. And A_1 achieved the target with other agents A_2 and A_3 lastly, then A_1 should distribute a part of reward got from environment to A_2 and A_3 as a group reward to praise them for their help. A_2 and A_3 would receive the reward from the agent A_1 who sent a new signal information but not from the agent who did not send new signal information. Because the cooperative action of agent A_1 and agent A_2, A_3 was strengthened as the learning process, when the signal information appears in the next time, the possibility of selecting the group cooperative behavior will rise. In Q-Learning method, Q-value is estimated on the basis of experience as follows:

Group Generation Reinforcement Learning

Initialize $Q_t(S_t^e \vee S_t^n, a_t)$ arbitrarily

Repeat (for each episode)

Initialize $S_t^e \vee S_t^n$

Repeat (for each step of episode)

Observe S_t^e

Cooperative communication of individual state

Receive S_t^n from other agents

Choose a_t from $S_t^e \vee S_t^n$ using policy derived from Q-value

Take action a_t

Observe r_t and next state $S_{t+1}^e \vee S_{t+1}^n$

Distribute r_t^G among agents

$$Q_t(S_t^e \vee S_t^n, a_t) \leftarrow Q_t(S_t^e \vee S_t^n, a_t) + \alpha[r_t + \beta \cdot r_t^G + \gamma \max_{a \in A} Q_t(S_{t+1}^e \vee S_{t+1}^n, a) - Q_t(S_t^e \vee S_t^n, a_t)]$$

$$S_t^e \vee S_t^n \leftarrow S_{t+1}^e \vee S_{t+1}^n$$

Until $S_t^e \vee S_t^n$ is terminal

where,

β : the rate of social interaction about group reward,

r_t^G : the group reward in current moment t .

The action evaluation value $w(s, b)$ is the same as Eq.(1) in Profit Sharing method, but reinforcement function $f(h)$ becomes as follows in Eq.(2).

$$f(h) = \alpha \cdot \gamma^h (r_t + \beta \cdot r_t^G) \quad (2)$$

For the agent of sending new signal information, it is necessary to supply a part of reward to the agents giving help and r_t^G is defined by $-\theta \cdot r_t$, and θ is the percentage of given reward as compared with r_t ($\theta \leq \theta_{lim}$). In GG-RL, the reward is re-distributed among agents, but the total reward value is not changed comparing with normal RL.

For the agents of giving cooperation, each agent will obtain a part of reward from the agent who sent signal as group reward, the reward of $\theta \cdot r_t$ is distributed by n which means the number of agents giving cooperative behavior to new signal sending agent ($n \leq n_{lim}$). Therefore, r_t^G is defined by $(\theta \cdot r_t)/n$.

In addition, the agents who regularly made the similar decision have an identity of cooperation like the social identity in our human society. Based on social identity theory, these agents create an in-group or self-categorization if they had stronger organization than agents of normal RL method. Based on group identity, we proposed group generation reinforcement learning (GG-RL) method. In GG-RL, the perfect action according with the original expected behavior will be praised through a group reward. Under the influence of group identity, group generation becomes possible.

Therefore, all agents not only receive rewards from the environment but also exchange the reward among agents to reinforce the coordinative behavior autonomously when the expected actions appeared. Agents are able to strength their group identity by taking reward exchanging as the social interaction strategy like human in society.

III. SIMULATION EXPERIMENTS

In order to verify the effectiveness and robustness of proposed methods, we also performed the comparison prey chase simulation experiments with obstacles using Profit Sharing method. This simulation is programmed by C language with the graphic library of OpenGL executed in Linux system. The view of simulator is shown in Fig.2.

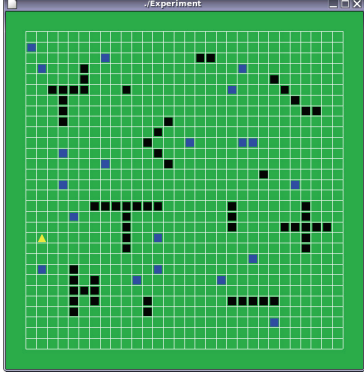


Fig. 2. The view of prey chase simulator of Exp.(1)

A. Simulation Preconditions

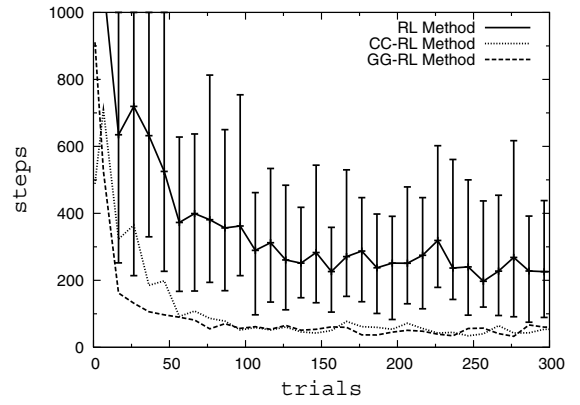
The preconditions of this simulation are set as follows:

- The field is composed by 30×30 grids in two dimensional environment. There are 62 obstacles in the learning environment.
- 20 hunter agents are set and represented as a square on the grid. The action and the perception ability of all agents are equal.
- Both the hunter and a prey represented as triangle agents can move in eight directions: up, down, left, right, up-right, up-left, down-right, down-left. They can stay at their current positions.
- The perception range is set 7×7 grids around itself.
- The default positions of all agents are arranged randomly and a prey agent moves always randomly.
- All hunter agents can obtain the other agents' perception information by sending new signal.
- When the prey agent becomes adjacent by three or more hunter agents, the task of chase is achieved. All the agents will return to the initial state after that.
- Hunter agent's task is to capture the prey agent as soon as possible. When the prey capturing is achieved, they will obtain the basic reward R assumed to be 100.

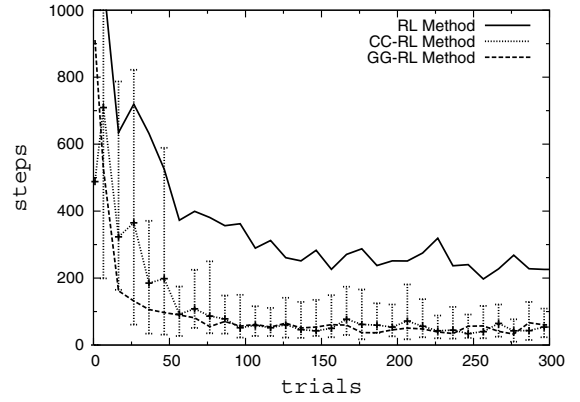
The parameters in this simulation are set as follows: $\gamma = 0.8$, $\lambda = 0.8$, $\alpha = 0.8$, $\beta = 1$, $\theta = 0.2$. To confirm the effectiveness of the proposed method, we performed comparison experiments with normal Profit Sharing reinforcement learning method. When an action is selected in Profit Sharing, the action is generally decided by the roulette selection of the action evaluation value added to the rule.

B. Simulation Results

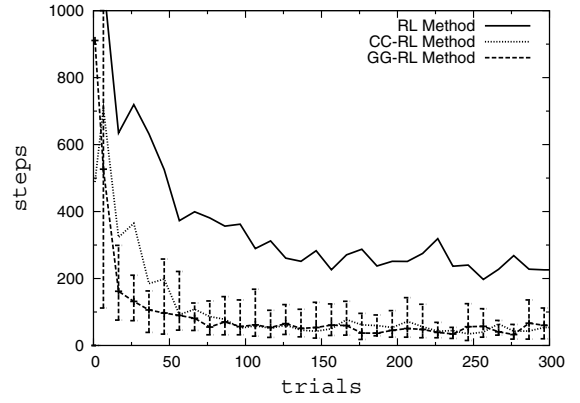
We performed two simulation experiments in different conditions: Exp.(1) with only one target; Exp.(2) with 5 targets. In



(a) Deviation of RL



(b) Deviation of CC-RL



(c) Deviation of GG-RL

Fig. 3. Results of Exp.(1) with one target

each experiment we compared four learning methods of RL, CC-RL and GG-RL. All of the methods are performed 5 times in this simulation, the average and deviation of simulation results are shown in result graphs.

Exp.(1) with only one target

In this experiment, there is only one target in the learning environment as shown in Fig.2. Vertical axis shows the steps required for capturing the prey and horizontal axis shows the trials of learning. From the average results in Fig.3(a)~(c), we confirmed that the cooperative communication reinforcement learning (CC-RL) method obtained better results than normal

RL method, the group generation reinforcement learning (GG-RL) method obtained the best learning result than the other methods. From the deviation results in Fig.3(a)~(c), we also confirmed that the stability of CC-RL and GG-RL methods are better than normal RL method. Furthermore, we can find that GG-RL method obtained better results than CC-RL method in the average and the deviation. That shows the exchanging group rewards among agents can reinforce the group behavior to achieve the task efficiently.

Exp.(2) with 5 targets

As a social agent, it is necessary to learn the best decision even with several targets in the complex environment. So we performed the simulation experiment with 5 targets as shown in Fig.4. The comparison experiment results of 3 methods are shown in Fig.5(a)~(c). By these results, we found that they are similar to the results of experiment with only one target. But the degree of convergence was not as good as Exp.(1) because it is more difficult naturally to capture 5 preys than only one.

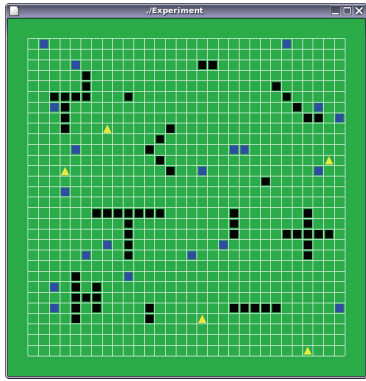


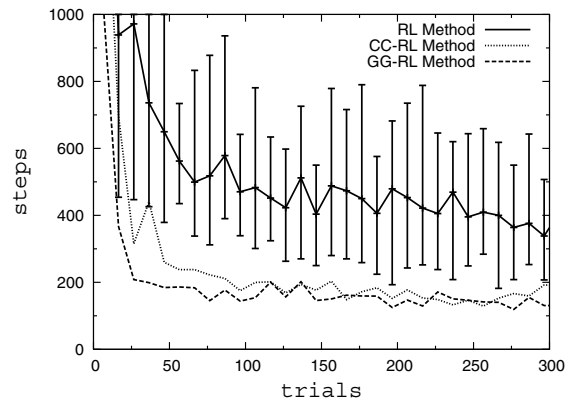
Fig. 4. Chase simulation of Exp.(2) with 5 targets

C. Individual and Group Growth

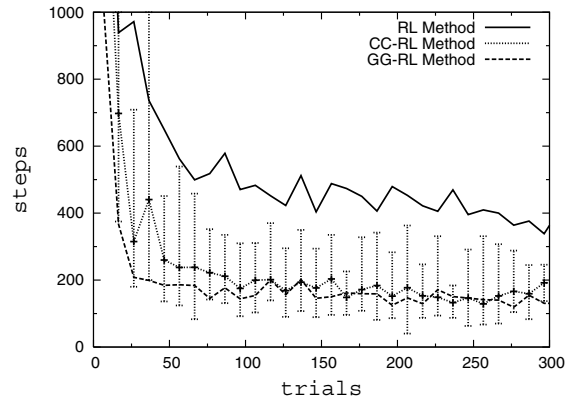
In order to distinguish the process of each individual growth, we also recorded the performance about 20 hunter agents. The growth of them is shown in Fig.6 based on catch time for preys. From this graph, we found that their performance is various after 300 trials to achieve the task. The best individual (agent 13) could get several times than the worst (agent 11) one in the achievement.

We have aimed at the generation of group behavior and recorded the interactive information among agents like Fig.7. The numbers of upside and left mean 20 agents and other numbers represent exchanging reward after 300 trials. The exchanging reward is only distributed among agent when the agents who regularly made the similar decision which means they have an identity of cooperation. Therefore, the exchanging reward represents the group identity of different agents.

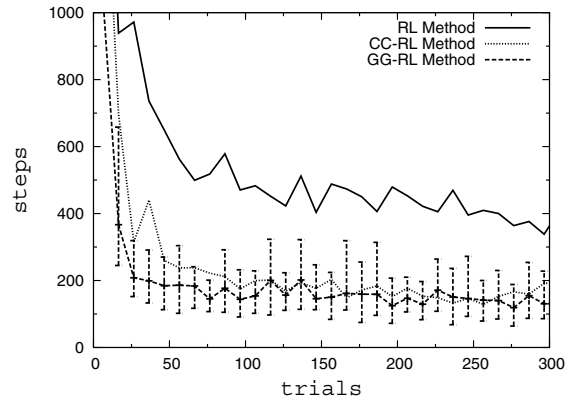
In the cooperative task, all the agents maybe chose anyone to cooperate, but the agent will more likely to cooperate with the agent who holds maximum exchanging reward. For each agent, if it holds maximum exchanging reward with many agents, it will have a greater influence in these agents. From the graph of Fig.7, agent 13 holds maximum exchanging reward



(a) Deviation of RL



(b) Deviation of CC-RL



(c) Deviation of GG-RL

Fig. 5. Results of Exp.(2) with 5 targets

with other 10 agents and we defined the pair influence of agent 13 as 10. All the agents will have a pair influence value based on this idea. Lastly we can easily get the relationship of all the agents like in Fig.8. The agent having greater pair influence is represented by wider circle area filled by color. The interactive degree of each others is represented by width of connecting lines, the wider the deeper social interactive ability. Therefore, we can easily find the relationship of all the agents in the pair influence aspect. In addition, we found that agent 13 did the best performance in individual aspect and also had the greatest influence in group generation.

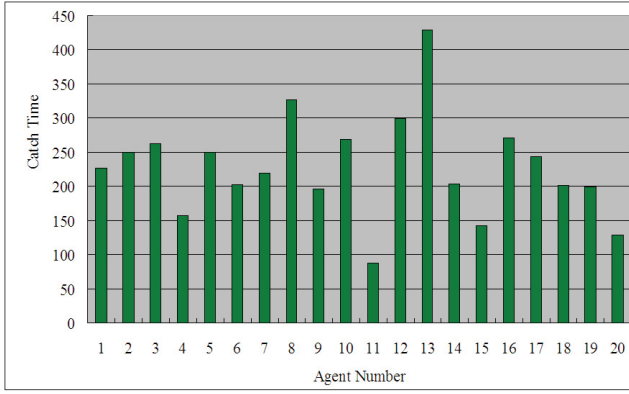


Fig. 6. Individual Growth

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0	350	350	160	180	190	120	430	110	170	100	220	300	130	110	130	120	150	100	120
2	350	0	320	150	170	180	190	400	190	200	80	220	400	170	90	330	100	110	110	70
3	350	320	0	150	270	160	180	200	170	170	60	220	370	180	140	310	150	140	120	60
4	160	150	150	0	140	130	160	80	130	220	60	110	70	100	120	120	140	110	90	20
5	180	170	270	140	0	270	220	210	380	160	80	260	350	110	140	180	200	200	220	110
6	190	180	160	130	270	0	200	230	80	350	100	240	190	50	60	110	160	150	100	130
7	120	190	180	160	220	200	0	290	130	240	30	210	400	70	70	140	130	140	130	130
8	430	400	200	80	210	230	290	0	140	230	100	310	570	310	100	350	370	140	120	60
9	110	190	170	130	380	80	130	140	0	170	50	130	210	100	90	90	100	60	160	90
10	170	200	170	220	160	350	240	230	170	0	40	440	330	140	140	130	110	240	140	80
11	100	80	60	60	80	100	30	100	50	40	0	110	120	50	20	70	50	40	20	20
12	220	220	220	110	260	240	210	310	130	440	110	0	510	110	100	150	230	120	190	90
13	300	400	370	70	350	190	400	570	210	330	120	510	0	450	140	380	400	120	150	80
14	130	170	180	100	110	50	70	310	100	140	50	110	450	0	40	170	200	120	90	40
15	110	90	140	120	140	60	70	100	90	140	20	100	140	40	0	120	30	90	80	50
16	130	330	310	120	180	110	140	350	90	130	70	150	380	170	120	0	150	70	70	40
17	120	100	150	140	200	160	130	370	100	110	50	230	400	200	30	150	0	100	90	60
18	150	110	140	110	200	150	140	140	60	240	40	120	120	120	90	70	100	0	140	40
19	100	110	120	90	220	100	130	120	160	140	20	190	150	90	80	70	90	140	0	40
20	120	70	60	20	110	130	130	60	90	80	20	90	80	40	50	40	60	40	40	0

Fig. 7. Group Identity Degree

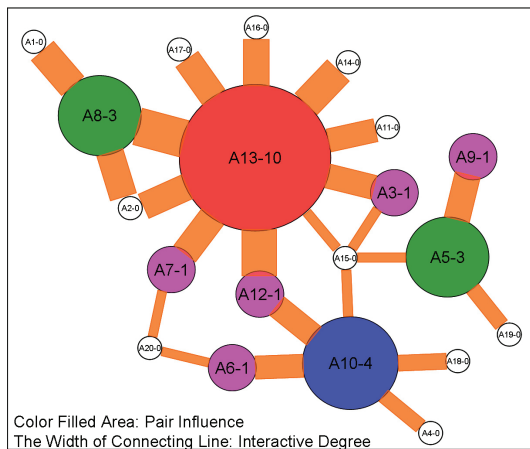


Fig. 8. Group Growth

IV. CONCLUSIONS

In this paper, we proposed CC-RL, GG-RL of multi-agent reinforcement learning methods with an excellent cooperative ability through the social interaction of cooperative communication and group generation by reward exchange which circulate among autonomous agents in MAS.

From the results of the comparison simulation between RL and CC-RL methods, we found that CC-RL method obtained the target state and individual information and could obtain the better cooperative ability than normal RL method. Although the outside of the perception range, the agent can obtain some environment states that can not be obtained from their restricted perception through the interaction with the other agents. The emergence of new undefined signal and the relative direction help the agent not only to perceive more learning states but also to get some information about the target in the task. Therefore, the communication of sending a new undefined signal accelerate their cooperative behavior.

Furthermore, we proposed GG-RL method that can give some evaluation to the action of agents by exchanging rewards. Rational re-distribution of reward causes the different learning results between CC-RL and GG-RL. Agents could encourage each other by exchanging reward. At last, the interaction among agents can be changed to the behavioral strategy communication like the human society and the autonomous cooperative group behavior of GG-RL method was best than the other methods.

At the same times, we can see the growth of 20 agents and group generation. The possibility of cooperation became very different from each other after learning. In the simulation, it is difficult to communicate more undefined signals, but it is possible to communicate and reinforce more complex behavior by proposed method. In the further works, we have a plan to add more communication modes to agents and let social agents have robust autonomous cooperative learning ability even in the competitive environment.

REFERENCES

- [1] M.J.Wooldridge, *An Introduction to MultiAgent Systems*, John Wiley and Sons, Ltd. England, 2002.
- [2] M.J.Mataric, "Reinforcement Learning in the Multi-Robot Domain," *Autonomous Robots*, 4, pp.73-83, 1997.
- [3] C.Claus, C.Boutillier, "The dynamics of reinforcement learning in cooperative multiagent systems," *AAAI/IAAI*, pp.746-752, 1998.
- [4] S.Arai, "Multiagent Reinforcement Learning Frameworks : Steps toward Practical Use," *Journal of The Japanese Society for Artificial Intelligence*, Vol.16, No.4, pp.476-481, 2001.
- [5] M.Tan, "Multi-Agent Reinforcement Learning: Independent vs. Cooperative Agents," *Proceedings of Tenth International Conference on Machine Learning*, pp.330-337, 1993.
- [6] R.Ribeiro, A.P.Borges, F.Enembreck, "Interaction Models for Multiagent Reinforcement Learning," *Computational Intelligence for Modelling, Control and Automation, International Conference*, pp.464-469, 2008.
- [7] M.Fasli, "Social Interactions in Multi-agent Systems: A Formal Approach," *2003 IEEE/WIC International Conference on Intelligent Agent Technology (IAT'03)*, pp.240-246, 2003.
- [8] K.Zhang, Y.Maeda, Y.Takahashi, "Group Behavior Learning in Multi-Agent Systems Based on Social Interaction among Agents," *Joint 5th International Conference on Soft Computing and Intelligent Systems and 11th International Symposium on advanced Intelligent Systems*, TH-B3-1, pp.193-198, 2010.