Learning of Soccer Player Agents Using a Policy Gradient Method: Pass Selection

Harukazu Igarashi, Hitoshi Fukuoka, and Seiji Ishihara

Abstract— **This research develops a learning method for the pass selection problem of midfielders in RoboCup Soccer Simulation games. A policy gradient method is applied as a learning method to solve this problem because it can easily represent the various heuristics of pass selection in a policy function. We implement the learning function in the midfielders' programs of a well-known team, UvA Trilearn Base 2003. Experimental results show that our method effectively achieves clever pass selection by midfielders in full games. Moreover, in this method's framework, dribbling is learned as a pass technique, in essence to and from the passer itself. It is also shown that the improvement in pass selection by our learning helps to make a team much stronger.**

Index Terms— **Multi-agent system, Pass selection, Policy gradient method, Reinforcement learning, RoboCup**

I. Introduction

Recently, much work has been done on the learning of coordination in multi-agent systems [1][2]. The RoboCup Simulation League 2D is recognized as a test bed for such research because there is no need to control real robots and one can focus on learning coordinative behaviors among players. However, multi-agent learning continues to suffer from several difficult problems such as state-space explosion, concurrent learning [3], incomplete perception [4], and credit assignment [2]. These four problems should be studied and resolved to make multi-agent learning successful in the games of the RoboCup Simulation League 2D (Fig.1).

As an example of multi-agent learning in a soccer game, Igarashi *et al*. proposed and applied a reinforcement-learning approach to realizing coordination play between a kicker and a receiver in direct free kicks [5]. They dealt with a learning problem between a kicker and a receiver when a direct free kick is awarded just outside the opponent's penalty area. In such a situation, to which point should the kicker kick the ball? They proposed a function that expresses heuristics to evaluate a candidate target point for effectively sending/receiving a pass and scoring. However, they dealt only with the attacking problems of 2v2 (two attackers and two defenders). In this paper, we apply the method to a pass selection problem of four midfielders in a full soccer game.

Manuscript received September 7, 2009. H. Igarashi and H. Fukuoka are with Shibaura Institute of Technology, 3-7-5 Toyosu, Koto-ku, Tokyo 135-8548, JAPAN (e-mail: arashi50@sic.shibaura-it.ac.jp).

S. Ishihara is with Kinki University, Hiroshima, JAPAN (e-mail: ishihara@hiro.kindai.ac.jp).

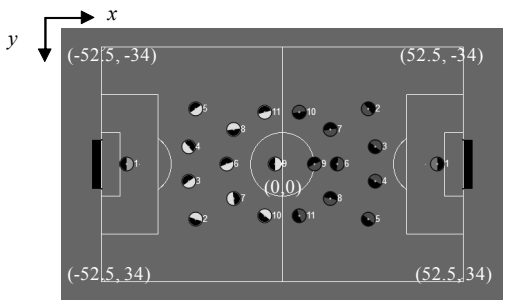


Fig.1. RoboCup Soccer Simulation League 2D.

II. Coordination of Soccer Agents

A. Cooperative Play in RoboCup Soccer Simulation

Reinforcement learning is widely used [6][7] in the research areas of multi-agent learning. In the RoboCup Soccer Simulation League, Andou used Kimura's stochastic gradient ascent (SGA) method to learn the dynamic home positions of 11 players [8]. Riedmiller *et al*. applied TD learning to learn such individual skills as intercepting the ball, going to a certain position, or kicking and to select those individual skills [9]. They investigated attacking problems with 2v1, 2v2, 3v4, and 7v8. Stone *et al*. studied keepaway problems with 3v2 [10] and half-field offense problems with 4v5 [11] using Sarsa [6] to learn the selection of macro behaviors such as ball holding, passing, dribbling, and shooting.

B. Coordination at Free Kicks

In the previous section, we cited several research efforts on the cooperative behaviors of soccer agents. However, a crucial problem remains. In the previous research, each agent apparently learns its policy of action selection "autonomously" to complete the given task. However, Riedmiller *et al*. assumed that all agents share input information, i.e., the x-y positions of all players and the ball, with other agents [9]. Stone *et al*. used other agents' experiences, which are time-series data on state, action, and reward, to accelerate learning in a large problem. For that purpose, agents must communicate their experiences with their partners to facilitate sharing information among them. If agents share input information and experiences with other agents, all agents will obtain the same value function by learning. This will simplify the realization of various cooperative plays among agents. However, if agents' observations are imperfect or uncertain, none of the agents can share the same input information with all other agents:

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Without perfect communication among agents, they cannot share their experiences with each other. Moreover, if only agents that have identical value functions are assumed, agent individuality and division of roles among them will not emerge from learning.

Igarashi *et al*. proposed a method where all agents learn autonomously without assuming perfect communication or identical input information [5]. They applied it to an attacking problem with 2v2 when a direct free kick is awarded just outside the opponent's penalty area. We briefly summarize the method in the next section.

III. Learning by a Policy Gradient Method

A. Policy Gradient Method

A policy gradient method is a kind of reinforcement learning scheme that originated from Williams's REINFORCE algorithm [12]. The method locally increases and maximizes the expected reward per episode by calculating the derivatives of the expected reward function of the parameters included in a stochastic policy function. This method, which has a firm mathematical basis, is easily applied to many learning problems. One can use it for learning problems even in non-Markov Decision Processes [13][14]. It was applied to pursuit problems where the policy function consists of state-action rules with weight coefficients that are parameters to be learned [14].

B. Stochastic Policy for Action Decision

Igarashi *et al*.[5] state policy $\pi(a;s,\omega)$ for determining action a ($\in A$) of all agents when a multi-agent system is in state s ($\in S$), and this policy is given stochastically by a Boltzmann distribution function with object function $E(a;s,\omega)$ as

$$\pi(a;s,\omega) \equiv \exp\left(-E(a;s,\omega)/T\right) \Big/ \sum_{x\in A}\exp\left(-E(x;s,\omega)/T\right) \quad (1)$$

Weight parameters $\omega=\{\omega_j\}$ (j=1,2,…,N_ω) in (1) are determined by a policy gradient method described in the next section. T is a parameter called *temperature*.

C. Autonomous Action Decision and Learning

For the autonomous action decisions and the learning of each agent, policy function $\pi(a;s,\omega)$ for the entire multi-agent system was approximated by the product of each agent's policy function $\pi_\lambda(a_\lambda;s_\lambda,\{\omega^\lambda_j\})$ as [13]-[15]

$$\pi(a;s,\omega) \approx \prod_\lambda \pi_\lambda\left(a_\lambda;s_\lambda,\left\{\omega^\lambda_j\right\}\right) , \quad (2)$$

where a_λ is the action of agent λ and s_λ is the state perceived by agent λ. ω^λ_j is the j-th parameter in agent λ's policy function $\pi_\lambda(a_\lambda;s_\lambda,\{\omega^\lambda_j\})$. In (2), it seems that the correlation among agent action decisions is neglected. However, each agent can see other agents' states, even if they are not perfect, and use them in its policy function $\pi_\lambda(a_\lambda;s,\{\omega^\lambda_j\})$. Thus, the approximation in (2) will partially contribute to learning of coordination among agents.

We assume that agent λ's policy function $\pi_\lambda(a_\lambda;s,\{\omega^\lambda_j\})$ is also given by a Boltzmann distribution function with objective function $E_\lambda(a_\lambda;s_\lambda,\{\omega^\lambda_j\})$ defined by

$$E_\lambda\left(a_\lambda;s_\lambda,\left\{\omega^\lambda_j\right\}\right) = -\sum_j \omega^\lambda_j \cdot U_j\left(a_\lambda;s_\lambda\right) , \quad (3)$$

where function $U_j(a_\lambda;s_\lambda)$ is the j-th heuristics that evaluates action a_λ [5].

At the end of each episode, common reward r is given to all agents that made their decisions during the episode. The derivative of expectation of reward $E[r]$ for parameter ω^λ_j is shown as

$$\partial E[r]\Big/\partial\omega^\lambda_j = E\left[r\sum_{t=0}^{L-1} e_{\omega^\lambda_j}(t)\right] , \quad (4)$$

where L is an episode length.

With (2), characteristic eligibility e_ω on the right-hand side of (4) can be written as [13][14]

$$e_{\omega^\lambda_j}(t) \equiv \partial \ln\pi\left(a(t);s(t),\left\{\omega^\lambda_j\right\}\right)\Big/\partial\omega^\lambda_j \approx \partial \ln\pi_\lambda\left(a_\lambda(t);s_\lambda(t),\left\{\omega^\lambda_j\right\}\right)\Big/\partial\omega^\lambda_j . \quad (5)$$

Substituting (2) and (3) into (5), the policy gradient in (4) gives a learning rule as

$$\Delta\omega^\lambda_j = \varepsilon r\sum_{t=0}^{L-1}\left[U_j\left(a_\lambda(t)\right) - \sum_{a_\lambda} U_j(a_\lambda)\pi_\lambda\left(a_\lambda;s_\lambda(t),\left\{\omega^\lambda_j\right\}\right)\right]\Big/T , \quad (6)$$

where $a_\lambda(t)$ is an action actually selected by policy π_λ and $s_\lambda(t)$ is a state perceived by agent λ at time t. Each agent updates ω^λ_j by the learning rule in (6) at the end of each episode [5].

IV. Pass Selection Problem

A. Pass Selection in a Soccer Game

A pass is a typical cooperative play between two players in a soccer game. Determining the receiver from among teammates is very important in a pass play. The first action to consider is passing the ball to a receiver safely. However, useless iteration of backward passes should be avoided. Thus, a player must select a receiver who stands in a position to receive the pass safely and who has a relatively high possibility of scoring a goal after receiving the ball. For this purpose, we use heuristics functions that seem to be useful for selecting a pass receiver as $U_j(a_\lambda;s_\lambda)$ in (3), where a_λ is a passer λ's action of selecting a pass receiver and passing the ball to the receiver.

B. Reward

In this paper, we apply the policy gradient method summarized in Section III to pass selection problems of midfielders. Dribbling is considered a pass from and to the passer itself. If midfielders can pass and dribble the ball

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

safely without being intercepted, the length of time their team holds the ball will be longer. Keeping the ball for as long a time as possible is obviously one of the effective strategies of midfielders in a soccer game. Here, we define the term "team *X* is keeping the ball" as the situation where the nearest player to the ball is a member of team *X* and this situation continues for a duration of more than five simulation cycles.

We define learning episode σ of team *X* by a history of states and actions of agents while team *X* keeps the ball, and we define reward *r* of episode σ as $r(\sigma) = -1/L(\sigma)$. Length $L(\sigma)$ of episode σ is defined by the difference between the time when team *X* begins to keep the ball and the time when the ball is taken by the opponent team. Since *r* takes a negative real value, it is actually a penalty rather than a reward. The shorter the duration time of keeping the ball is, a larger common penalty $r(\sigma)$ is given to team *X*'s midfielders who made pass selection during episode σ.

C. Heuristics for a Pass Selection Problem

We used five heuristics from U_1 to U_5 for evaluating the current position of a teammate as a desirable pass receiver. U_1, U_2 and U_3 are heuristics for passing the ball safely. U_4 is heuristics for making an aggressive pass. U_5 is used for treating reliable information as more important knowledge than unreliable information. All U_i are normalized between 0 and 10.

The meanings of the five heuristics are summarized as follows. U_1 considers the existence of opponents on the pass course. U_2 evaluates a distance between a receiver and the opponent nearest to the receiver. U_3 takes into account the number of opponent players around a receiver. U_4 expresses a heuristics that the nearer receiver to the opponent's goal has a greater chance of scoring a goal. U_5 evaluates the degree of certainty of a receiver's position perceived by the passer. The definitions of the five heuristics are shown in Appendix.

V. Experiments

A. Team Used in Learning Experiments

We used UvA Trilearn Base 2003 as a base team for learning experiments. UvA Trilearn 2003 is a team of the University of Amsterdam and champion over 46 qualified teams in the 2D Simulation Soccer League of RoboCup2003 Padova. The team released a part of the source code of UvA Trilearn 2003 [16].

According to the home page of the UvA team [16], the released code contains low-level and intermediate-level implementation (agent-environment synchronization method, world model, player skills) but not a high-level decision procedure. Instead, they have included a simple high-level action selection strategy. The fastest player to the ball intercepts the ball and shoots it to a random corner in the opponent's goal regardless of his position on the field. The remaining players move to a strategic position determined by their home position in the formation and by the position of the ball. We modified the code slightly and implemented the learning function defined in Sections III and IV for UvA's four midfielders. We call this team "UvA Trilearn Base 2003," or "UvA Base" for short, and use it in our learning experiments.

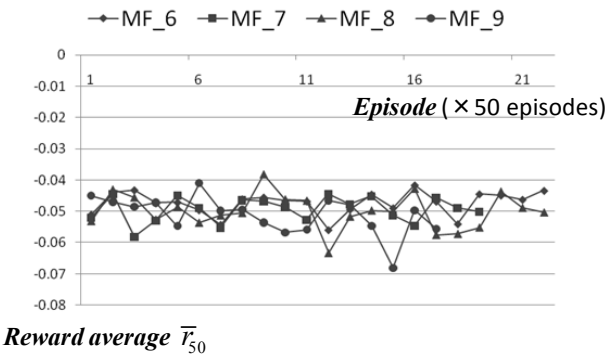


Fig. 2. Reward average over the last 50 episodes.

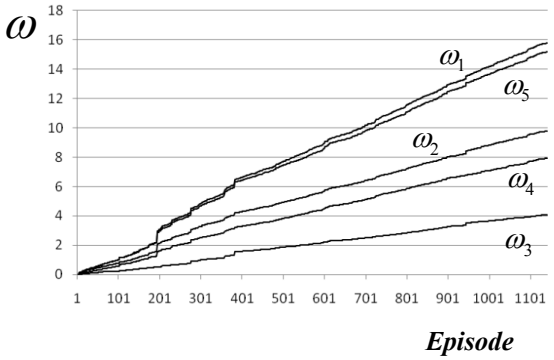


Fig. 3. Change in ω_j of player MF_8.

B. Learning Experiments

UvA Base with four midfielders who learn pass selection plays 50 games against UvA Base with midfielders who randomly make passes to teammates. Initial values of ω^{λ}_j in (3) are set to 0. *T* and ε in (6) are set to 10.0 and 0.1, respectively.

The change in rewards *r* given to four midfielders is shown in Fig.2, where $\overline{r_{50}}$ is an average over the last 50 episodes. Midfielders have their own set of ω^{λ}_j and learn their values by playing 50 games. The results of ω^{λ}_j are nearly identical to each other. Therefore, only the results of a midfielder called MF_8 are shown in Fig. 3.

Reward averages $\overline{r_{50}}$ for all midfielders do not increase in Fig. 2, even as the learning proceeds. At a glance, the learning seems ineffective and a waste of time. However, the effectiveness of learning is shown by evaluation experiments in the next section. The phenomenon of reward averages not increasing is due to the following reason. A player in UvA Base is programmed to keep a close watch over the opponent goal when it possesses the ball. This increases confidence in the information of a teammate's position in the forward direction of a passer, and it increases the importance of a forward pass in objective function $E(a_\lambda;s,\{\omega^{\lambda}_j\})$ in (3). Thus, a large value of ω_5 in Fig. 3 causes an agent to make a more offensive pass and to play in a deeper area of the opponent team's territory as learning proceeds. Playing in a deep area of the opponent prevents the duration of keeping the ball from getting longer.

The ratio among ω^{λ}_j seems to converge in Fig.3. This means that there is a strong attractor point in the parameter

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

space of ω^{λ}_{j} and that our learning captures this attractor point. Increasing ω^{λ}_{j} with a fixed ratio in magnitude is equivalent to annealing in a stochastic policy in (1). Accordingly, a deterministic pass selection algorithm based on ω^{λ}_{j} at the attractor point was obtained as policy $\pi_{\lambda}(a_{\lambda};s,\{\omega^{\lambda}_{j}\})$ by the learning experiments in this section.

C. Evaluation Experiments

In order to evaluate the effectiveness of the learning method described in Sections III and IV, we have the team of UvA Base with four midfielders trained in the learning experiments play 30 games against UvA Base with midfielders who did not obtain any learning. The four midfielders in the former team make a deterministic decision because parameter T in their policies is fixed at a very low value in the evaluation experiments. The midfielders in the latter team randomly make passes to teammates because all ω^{λ}_{j} are fixed to 0.

Table 1 shows that learning pass selection in only five games(N=5) increases the team's goals and decreases the opponent's goals. Therefore, passes by the four midfielders contribute to both offense and defense. Furthermore, it is occasionally observed in full games that a midfielder selects himself as a receiver and begins dribbling when there is no appropriate receiver to whom the ball can be safely passed. Consequently, they learned how to make a decision on whether they should select passing or dribbling the ball.

Table 1. Results of 30 games between UvA Base with four midfielders who learned pass selection in N games and UvA Base with midfielders who did not learn but only randomly made passes to teammates.

N	Games	Wins -draws-losses	Goals
0	30	14-5-11	31-24
5	30	22-2-6	59-23
10	30	23-2-5	51-17
30	30	23-1-6	46-15
50	30	26-0-4	54-19

VI. Conclusion

This research develops a learning method for the pass selection problem of midfielders in RoboCup Soccer Simulation games. A policy gradient method is applied as a learning method to solve this problem because it can easily express the various heuristics of pass selection in a policy function. We implement the learning function in the midfielders' programs of a well-known team, UvA Trilearn Base 2003. Experimental results show that our method effectively achieves clever pass selection by midfielders in full games. Moreover, in this method's framework, dribbling is learned as a pass technique, in essence to and from the passer itself. It is also shown that the improvement in pass selection by our learning helps to make a team much stronger.

In the future, we will apply our learning method to other tasks in a soccer game, such as a receiver's selection of the destination point for receiving a pass. A receiver must move to receive a pass safely, to receive a through pass from a passer, and to deceive opponent players. Consequently, the receiver must learn some policies and change them depending on the situation. Moreover, an agent becomes a passer and a receiver depending on the time and situation. For example, a passer and a receiver must change their roles at the next moment to succeed with a wall-pass. Accordingly, an agent must change its role and policy at every moment. The next step of our work will focus on how agents learn the most desirable selection of roles and policies in their current situation.

Appendix

We define five heuristics from U_1 to U_5 for evaluating the current position of a teammate as a desirable pass receiver. All U_i are normalized between 0 and 10.

(i) $U_1(a_{\lambda};s_{\lambda})$ considers the existence of opponents on the pass course and is defined by

$$U_1 = \begin{cases} 10.0\ (if\ diff > 30\ or\ OppDist > AgentDist \\ \qquad or\ \text{when dribbling}) \\ 9.0\ (if\ 20 < diff \le 30 \,\&\, OppDist < AgentDist) \\ 7.0\ (if\ 10 < diff \le 20 \,\&\, OppDist < AgentDist) \\ 4.0\ (if\ diff \le 10 \,\&\, OppDist < AgentDist) \\ 2.0 \quad (else) \end{cases}, \qquad (7)$$

where *diff* is the angle between the pass direction and the direction to an opponent. *AgentDist* is the distance between a passer and a receiver. *OppDist* is the distance between a passer and the nearest opponent player to the passer (Fig.4a). If a passer cannot get enough information to calculate all three values, we set U_1=2.0. If a passer perceives more than one opponent, values of U_1 are calculated for all opponents and the smallest of these values is used.

(ii) $U_2(a_{\lambda};s_{\lambda})$ evaluates distance *min_dist* between a receiver and the opponent nearest to the receiver (Fig.4b). It is defined by

$$U_2 = \begin{cases} 10 & (if\ \text{min\_dist} \ge 30) \\ 10 - (30.0 - \text{min\_dist}) / 5.0 & (if\ \text{min\_dist} < 30) \\ 2.0 & (else) \end{cases}. \qquad (8)$$

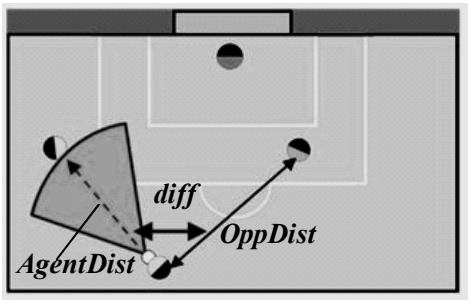
(iii) $U_3(a_{\lambda};s_{\lambda})$ takes into account the number of opponent players around a receiver. *OpponentsNum* is the number of opponent players standing within a 10-meter radius of the receiver:

$$U_3 = \begin{cases} 10 - OpponentsNum\ (if\ OpponentsNum \le 10) \\ 0.0\ (else) \end{cases}. \qquad (9)$$

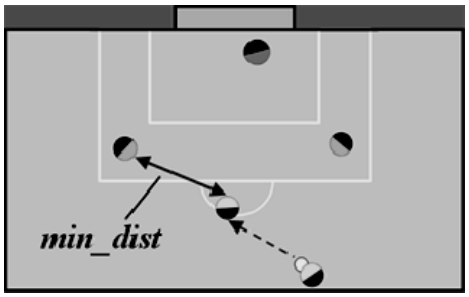
(iv) $U_4(a_{\lambda};s_{\lambda})$ expresses a heuristics that the nearer receiver to the opponent's goal has a greater chance of scoring a goal. It is defined by

$$U_4 = \begin{cases} 10 - goal\_dist / 10.0\ (if\ goal\_dist \le 40.0) \\ 2.0 \qquad (else) \end{cases}, \qquad (10)$$

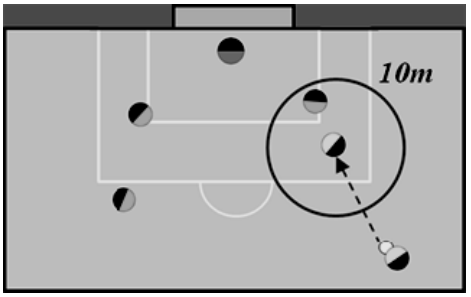
ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)



(a) U_1



(b) U_2



(c) U_3

(d) U_4

Fig. 4. Heuristics used for a passer to select a receiver.

where *goal_dist* is the distance between a receiver and the goalmouth of the opponent team.

(v) $U_5(a_\lambda;s_\lambda)$ evaluates the degree of certainty of a receiver's position perceived by the passer and is defined by

$$U_5 = \begin{cases} 9.0 \ (\textit{if} \ \text{a receiver and a passer is the same player}) \\ 10.0 \times \textit{confidence} \quad (\textit{else}) \end{cases}, \quad (11)$$

where *confidence* is the reliability of the receiver's position. The value *confidence* is defined by

$$\textit{confidence} \equiv \max\left(1 - \frac{\textit{CurrentCycle} - \textit{LastSeeCycle}}{100}, 0\right), \quad (12)$$

where *CurrentClcye* is the current time and *LastSeeCycle* is the latest time when the passer saw a receiver. The later the information is, the higher its reliability is.

REFERENCES

[1] G. Weiss, S. Sen, Eds. *Adaption and Learning in Multi-agent System*. Germany : Springer-Verlag, 1996.

[2] S. Sen, G. Weiss, "Learning in Multiagent Systems", in *Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence*, G. Weiss, Ed. The MIT Press, 1999, pp. 259-208.

[3] S. Arai, and K. Miyazaki, "Learning Robust Policies for Uncertain and Stochastic Multi-agent Domains," in *7th International Symposium on Artificial Life and Robotics*, 2002, pp. 179-182.

[4] W. S. Lovejoy, "A Survey of Algorithmic Methods for Partially Observed Markov Decision Processes". *Annals of Operations Research*, vol. 28, 1991, pp. 47-66.

[5] H. Igarashi, K. Nakamura, and S. Ishihara, "Learning of Soccer Player Agents Using a Policy Gradient Method: Coordination between Kicker and Receiver during Free Kicks", in *2008 International Joint Conference on Neural Networks* (*IJCNN 2008*), 2008, pp. 46-52 (Paper No. NN0040).

[6] R. S. Sutton, and A. G. Barto, *Reinforcement Learning*. The MIT Press 1998.

[7] L. P. Kaelbling, M. L. Littman, and A.W. Moore, "Reinforcement Learning: A Survey", in *J. Artificial Intelligence Research*, vol. 4, 1996, pp. 237- 285.

[8] T. Andou, "Refinement of Soccer Agents' Positions Using Reinforcement Learning", in *RoboCup-97: Robot Soccer World Cup I* , H. Kitano, Ed. Berlin: Springer-Verlag, 1998, pp. 373-388.

[9] M. Riedmiller, and T. Gabel, "On Experiences in a Complex and Competitive Gaming Domain -Reinforcement Learning Meets RoboCup-", in *The 2007 IEEE Symposium on Computational Intelligence and Games* (*CIG2007*), 2007, pp. 17-23.

[10] P. Stone, G. Kuhlmann, M. E. Taylor, and Y. Liu, "Keepaway Soccer: From Machine Learning Test bed to Benchmark", in *RoboCup 2005: Robot Soccer World Cup IX*, A. Bredenfeld, A. Jacoff, I. Noda, Y. Takahashi, Eds. New York : Springer-Verlag, 2006, pp. 93-105.

[11] S. Kalyanakrishnan, Y. Liu, and P. Stone, "Half Field Offense in RoboCup Soccer –A Multiagent Reinforcement Learning Case Study", in *RoboCup-2006: Robot Soccer World Cup X*, G. Lakemeyer, E. Sklar, D. G. Sorrenti, and T. Takahashi, Eds., Berlin: Springer-Verlag, 2007, pp. 72-85.

[12] R. J. Williams, "Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning," *Machine Learning*, vol. 8, 1992, pp. 229-256.

[13] H. Igarashi, S. Ishihara, and M. Kimura, "Reinforcement Learning in Non-Markov Decision Processes -Statistical Properties of Characteristic Eligibility-," *IEICE Transactions on Information and Systems*, vol. J90-D, no. 9, 2007, pp. 2271-2280 (in Japanese). This paper is translated into English and included in *The Research Reports of Shibaura Institute of Technology, Natural Sciences and Engineering,* vol. 52, no. 2, 2008, pp. 1-7. ISSN 0386-3115

[14] S. Ishihara, and H. Igarashi, "Applying the Policy Gradient Method to Behavior Learning in Multiagent Systems: The Pursuit Problem", *Systems and Computers in Japan*, vol. 37, no. 10, 2006, pp. 101-109.

[15] L. Peshkin, K. E. Kim, N. Meuleau, and L. P. Kaelbling, "Learning to Cooperate via Policy Search", in *Proc. of 16th Conference on Uncertainty in Artificial Intelligence* (*UAI2000*), 2000, pp. 489-496.

[16] http://staff.science.uva.nl/~jellekok/robocup/2003/

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)