Multi-agent Reinforcement Learning: A Modular Approach

Norihiko ONO and **Kenji FUKUMOTO**
Department of Information Science and Intelligent Systems
Faculty of Engineering, University of Tokushima
2-1 Minami-Josanjima, Tokushima 770, Japan
ono@is.tokushima-u.ac.jp

Abstract

To investigate the potentials and limitations of multi-agent reinforcement learning, several attempts have been made to let multiple monolithic reinforcement-learning agents synthesize coordinated decision policies needed to accomplish their common goals effectively. Most of these straightforward reinforcement-learning approaches, however, scale poorly to more complex multi-agent learning problems because the state space for each learning agent grows exponentially in the number of its partner agents engaged in the joint task. In this paper, we consider a modified version of the Pursuit Problem as a multi-agent learning problem which is computationally intractable by those straightforward approaches. We show how successfully a collection of modular Q-learning hunter agents synthesize coordinated decision policies needed to capture a randomly-fleeing prey agent, by specializing their individual functionality and synthesizing herding behavior.

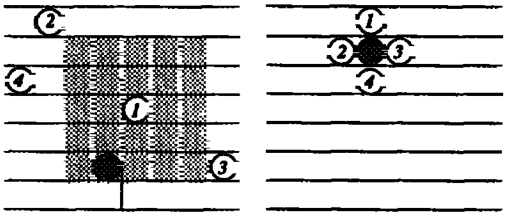
Introduction

In attempting to let artificial organisms or simple reactive robots synthesize some coordinated behavior, several researchers in the fields of artificial life and robotics have applied monolithic reinforcement-learning algorithms to multi-agent learning problems(e.g., (Drogoul 1991; Ono & Rahmani 1993; Ono, Ohira & Rahmani 1995a; Ono, Ohira & Rahmani 1995b; Rahmani & Ono 1993; Tan 1993; Yanco & Stein 1992)). In most of these applications, only a small number of learning agents are engaged in their joint tasks and accordingly the state space for each agent is relatively small. This is the reason why monolithic reinforcement-learning algorithms have been successfully applied to these multi-agent learning problems. However, these straightforward applications of reinforcement-learning algorithms do not successfully scale up to more complex multi-agent learning problems, where not a few learning agents are engaged in some coordinated tasks(Tan 1993). In such a multi-agent problem domain, agents should appropriately behave according to not only sensory information produced by the physical environment itself but also that produced by other agents, and hence the state space for each reinforcement-learning agent grows exponentially in the number of agents operating in the same environment. Even simple multi-agent learning problems are computationally intractable by the monolithic reinforcement-learning approaches. We consider a modified version of the Pursuit Problem(Benda, Jagannathan & Dodhiawalla 1985) as such a multi-agent learning problem, and show how successfully modular Q-learning prey-pursuing agents synthesize coordinated decision policies needed to capture a randomly-moving prey agent[1].

Problem Domain

In this paper, we treat a modified version of the Pursuit Problem(Benda, Jagannathan & Dodhiawalla 1985). In an $n \times n$ toroidal grid world, a single *prey* and four *hunter* agents are placed at random positions in the grid, as shown in Fig. 1 (a). The hunters operate attempting to capture the randomly-fleeing prey. At every time step, the prey and each hunter select their own actions independently without communicating with each other and accordingly perform them. Each agent has a repertoire of five actions. It can move in one of four principle directions (north, east, south, or west), or alternatively remain at the current position. The prey can not occupy the same position that a hunter does. However, more than one hunter can share the same position. So the prey can not move to a position which has been occupied by a hunter, and vice versa. A hunter has a limited visual field of depth d ($2d + 1 < n$) and has a unique identifier. The prey also has a unique identifier. A hunter accurately locates the relative position and recognizes the identifier of any other agent operating within its visual field, and it determines its next action according to the perceived information. The prey is captured when all of its four neighbor positions are occupied by the hunters, as shown in Fig. 1 (b). Then all of the prey and hunter agents move to new random positions in

[1] Preliminary results of this work have been reported in (Ono, Ikeda & Rahmani 1995).



(a) An initial state. (b) A goal state.

Figure 1: The pursuit problem. Four hunters and a prey are indicated by white and black circles, respectively. Each hunter is assigned an identifier (1,2,3, or 4) which is visible to the other hunters. A hunter is able to accurately recognize the relative position and the identifier of any other agent in its visual field of depth d. In the figure, when $d = 2$, the visual field of the hunter 1 is indicated by a shaded square in (a). The relative position of the prey to the hunter 1 is $(-1, -2)$ in this initial configuration.

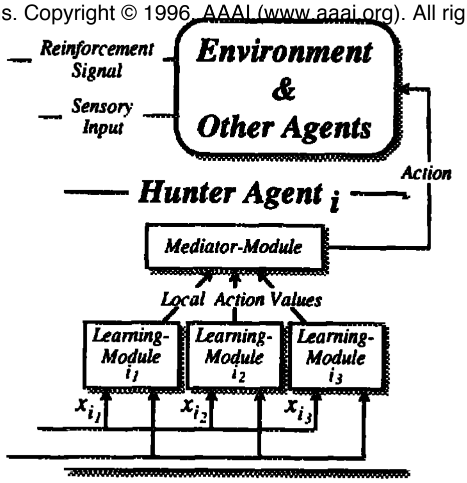


Figure 2: Modular architecture for a hunter agent. Learning-Module i_k only locates the partner i_k and the prey, and does not concern any other partners' positions.

the grid world and the next trial starts. The ultimate goal for the hunters is to capture the prey as frequently as possible.

This problem has been treated as a canonical problem in the field of distributed artificial intelligence (DAI), but never as a multi-agent reinforcement-learning problem. Researchers in DAI have attempted to suggest some coordinated strategies, based on their proposed distributed problem solving principles, and have evaluated the performance(e.g., (Gasser et al. 1989; Levy & Rosenschein 1991)). In this paper, we attempt to let reinforcement-learning hunter agents spontaneously synthesize some coordinated strategies. As a basic learning algorithm, we have chosen to employ Q-learning. However, the straightforward application of monolithic Q-learning to this kind of multi-agent learning problems brings about the problem of combinatorial explosion in the state space. Since multiple learning agents, each observing its partners' behavior, have to jointly perform a coordinated task, the state space for each learning agent grows exponentially in the number of its partners. For example, let us represent the state of a hunter agent by a combination of the relative positions of other agents, where the positions are represented by some unique symbols when they are not located within the hunter's visual field. Although there are only a small number of collaborative agents involved, the size of state space for a hunter, $((2d+1)^2+1)^4$, is enormous even when the visual field depth d is small; e.g., at $d = 2$ and $d = 3$, it amounts to $456,976$ and $6,250,000$ respectively.

For this reason, we have decided to construct each hunter agent by a variant of Whitehead's modular Q-learning architecture(Whitehead et al. 1993) as shown in Fig. 2. This architecture consists of three learning-modules and a single mediator-module. Each learning-module focuses on specific attributes of the current state and performs Q-learning. More specifically, each module only observes the relative position of the prey and that of a partner associated with some specific identifier and ignores any information concerning the other partners. Its state is represented by a combination of these relative positions and its size equals $((2d+1)^2+1)^2$. This size of state space is tractable by monolithic Q-learning algorithm so long as d is relatively small. On the other hand, a mediator-module, without learning, combines learning-modules' decision policies using a simple heuristic decision procedure and determines a final decision by the corresponding agent. In our preliminary experiments, the mediator-module selects the following action:

$$arg\max_{a\in A}\sum_{k=1}^{3} Q^{i_k}(x_{i_k},a)$$

where Q^{i_k} denotes the action-value function implemented by the learning-module i_k. This corresponds to what is called *greatest mass* merging strategy(Whitehead et al. 1993). Intuitively, this is a selection by majority among actions proposed by individual learning-modules[2].

Our modular Q-learning architecture is based on the original one proposed in (Whitehead et al. 1993), but they are not identical. The architecture in (Whitehead

[2] Previously we considered another variant of the pursuit problem where each hunter's identifier is *invisible* to any other agents. We used yet another modular approach to the problem, and the preliminary results are reported in (Ono, Fukumoto & Ikeda 1996).

et al. 1993) was particularly designed for constructing a *single* reactive planner pursuing a fixed set of dynamically-activated multiple goals. It was primarily employed for reducing the planner's enormous state space caused by the existence of multiple goals. Here, our architecture is designed for reducing each agent's intractably enormous state space caused by the existence of its partners jointly working in the same environment. Note also that all the learning-modules, in our modular architecture, are always active and participate in the reinforcement-learning task unlike in the Whitehead's case.

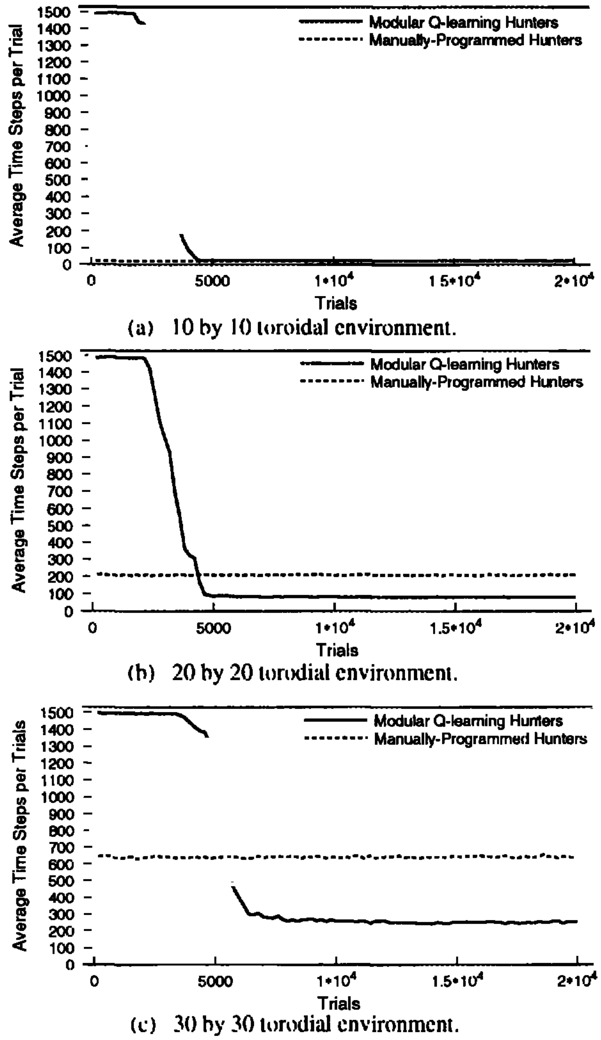
Simulation Results

Our simulation run consists of a series of trials, each of which begins with a single prey and four hunter agents placed at random positions, and ends when the prey is captured. A trial is aborted when the prey is not captured within 1,500 time steps. Upon capturing the prey, individual hunters immediately receive a reward of 1.0, and accordingly all of its constituent learning-modules uniformly receive the same reward regardless of what actions they have just proposed. Our hunters receive a reward of −0.1 in any other case. The learning rate α and discount factor γ are set to 0.1 and 0.9, respectively. Initial Q-values for individual state-action pairs are randomly selected from the interval $[0.01, 0.1]$.

The results of our simulation runs are shown in Fig. 3. The solid curves in this figure indicate the performance by the modular Q-learning hunters. At initial trials, the hunters can not capture the prey effectively, but shortly they start improving their performance dramatically. Eventually they come to steadily accomplish their task at every trial in any of our simulation runs. The meanings of dashed curves will be explained later.

To see the efficacy of our modular Q-learning approach compared with a straightforward monolithic version, we tried to solve the pursuit problem using a monolithic Q-learning approach. However, it turned out to be difficult to implement each hunter by a monolithic Q-learning even when each hunter's visual field depth is small, say 3. So we have to set the visual field depth to 2 or 1 in order to show the monolithic Q-learning hunters' learning performance. In Fig. 4, the performance by the implemented monolithic Q-learning hunters and that by the modular Q-learning hunters are shown. It may be possible to improve the monolithic Q-learning hunters' performance by fine tuning of the learning parameters, but limited computational experience showed that any significant improvement is difficult.

We have supposed that the prey selects its actions randomly. What will happen if the prey attempts to escape from the located hunters? It is easy to let such a prey totally outsmart our learning hunters, if it has the same visual field depth as that of a hunter. For ex-



(a) 10 by 10 toroidal environment.

(b) 20 by 20 torodial environment.

(c) 30 by 30 torodial environment.

Figure 3: Performance of modular Q-learning hunter agents. Four hunters, each with a fixed visual field depth 3, are operating together in environments with various dimensions. Solid curves indicate the performance by modular Q-learning hunters, while dashed curves indicate the performance by manually-programmed hunters, which as explained later, attempt to independently locate the prey by moving randomly and then behave in an optimal way as soon as locating it. Each curve is an average over 25 simulation runs. During the course of dramatically improving their performance, the learning hunters are specializing their individual functionality as explain below. Note that the larger the environment, the more clearly the learning hunters outperform the manually-programmed ones. The learning hunters outperform the programmed ones mainly by synthesizing a kind of herding behavior as illustrated below.

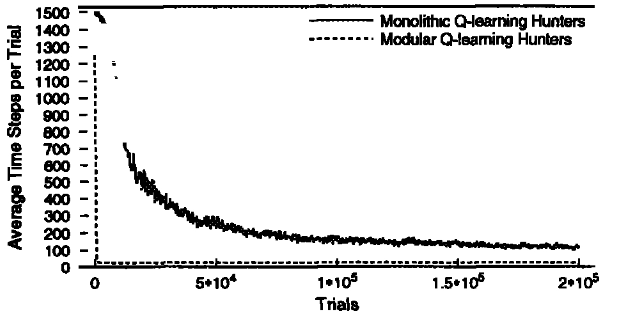


Figure 4: Performance of monolithic Q-learning hunter agents. A typical performance by monolithic Q-learning hunters is shown as well as that by the modular Q-learning ones. In either case, each hunter's visual field depth equals 2 and a 10×10 environment is employed. The same learning rate, discount factor, and Q-value initialization scheme as those described above are employed here.

ample, if the prey behaves randomly when no hunter is located, and otherwise it selects its actions so as to maximize the total sum of Manhattan-distances to the located hunters, then our learning hunters can never capture the prey. Usually, in this case the prey can escape from the located hunters very easily, and accordingly it is almost impossible for our learning hunters to get the first positive reward.

However, if the prey's visual field is shallower than that of the learning hunters, it is possible to let the learning hunters synthesize effective coordinated behavior to capture it, even when the prey moves using the moving strategy mentioned above. Figure 5 shows the typical results of simulation runs where the prey and each hunter have a visual field depth of 2 and 3, respectively. In these cases, as shown in the figure, the learning hunters synthesized some coordinated behavior which allows them to steadily capture the prey.

Synthesized Coordinated Behavior

Our modular Q-learning hunter agents exhibited two kinds of collective behavior, namely, *herding* and *functionality-specialization*. During the course of dramatically improving their performance, individual hunters are specializing their functionality. To capture the prey, individual hunters have to exclusively occupy one of four positions surrounding it. Let us call these positions *capture positions* of corresponding hunters. Eventually, our modular Q-learning hunters attempt to improve their pursuit performance by completely fixing their individual capture positions, i.e., by specializing their functionality. Figure 6 illustrates how such a functionality-specialization is typically developed among hunters where a 20×20 environment is used. Each of these four 3D bar graphs shows a

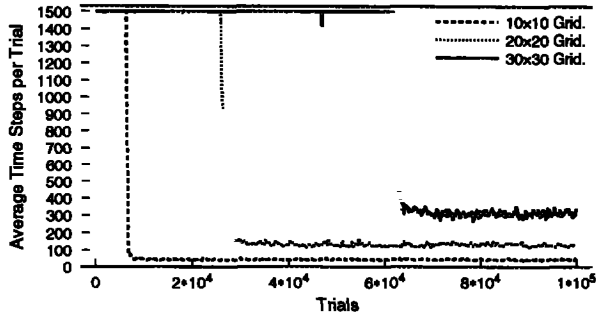


Figure 5: Performance of learning hunters pursuing an escaping prey. In any of these three cases, the hunters and the prey have a visual field depth of 3 and 2, respectively.

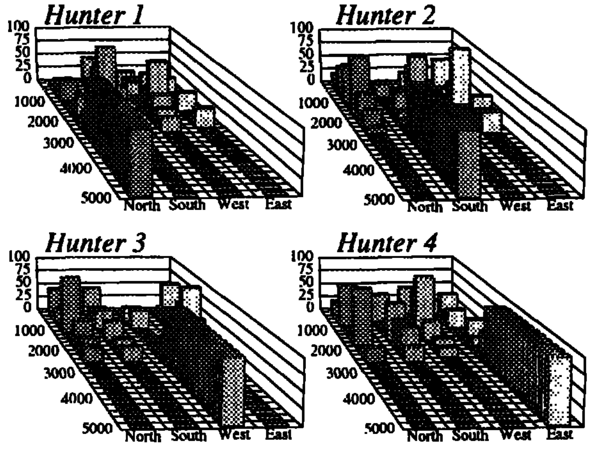


Figure 6: Typical specialization process by hunters.

transition of the probability distribution over capture positions (indicated by relative directions to the prey) taken by the corresponding hunter. At initial 2,900 trials, hunters have not fixed their own capture positions, but after trial 3,000, they have chosen their own capture positions and during this change in their behavioral patterns, their overall performance was dramatically improved.

Besides specializing their functionality, our hunters obtain a kind of herding behavior. Even if establishing the above-mentioned static role-assignment, hunters can not always accomplish the capturing task effectively until they have obtained an effective way of collectively locating the prey, since each of them can only observe a limited portion of the grid world. To see this, we manually-programmed a set of four hunters. We assigned individually these hunters' capture positions in advance, and then let them (i) attempt to independently locate the prey by moving randomly and (ii) reach their capture positions in a probabilistically optimal way as soon as locating it. The dashed curves in Fig. 3 show the performance by the programmed hunters. Note that the larger environment, the more clearly our learning hunters outperform these manually-programmed hunters.

Our learning hunters do not attempt to locate the prey independently, because it is not an effective way even when the prey moves randomly. Instead they locate it in a coordinated manner. Once having located a partner, a hunter attempts to keep the partner in sight so long as the prey has not been located and thereby it attempts to make a herd together with a partner. By herding in this way, hunters can operate as each other's eyes and accordingly compensate their limited perceptual capabilities. However, their behavior suddenly changes once one of herding members has happened to locate the prey. The member quickly attempts to reach its capture position and the other ones follow it and accordingly are likely to locate the prey. When all of hunters have located the prey, they can easily capture it.

Typically, the prey capturing process by our modular Q-learning hunters consists of two phases: (i) a *collective search phase* where all of hunters attempted to identify the prey's position by configuring a herd, and (ii) a *surrounding phase* where each hunter having identified the prey's position attempts to reach its fixed capture position in an effective way.

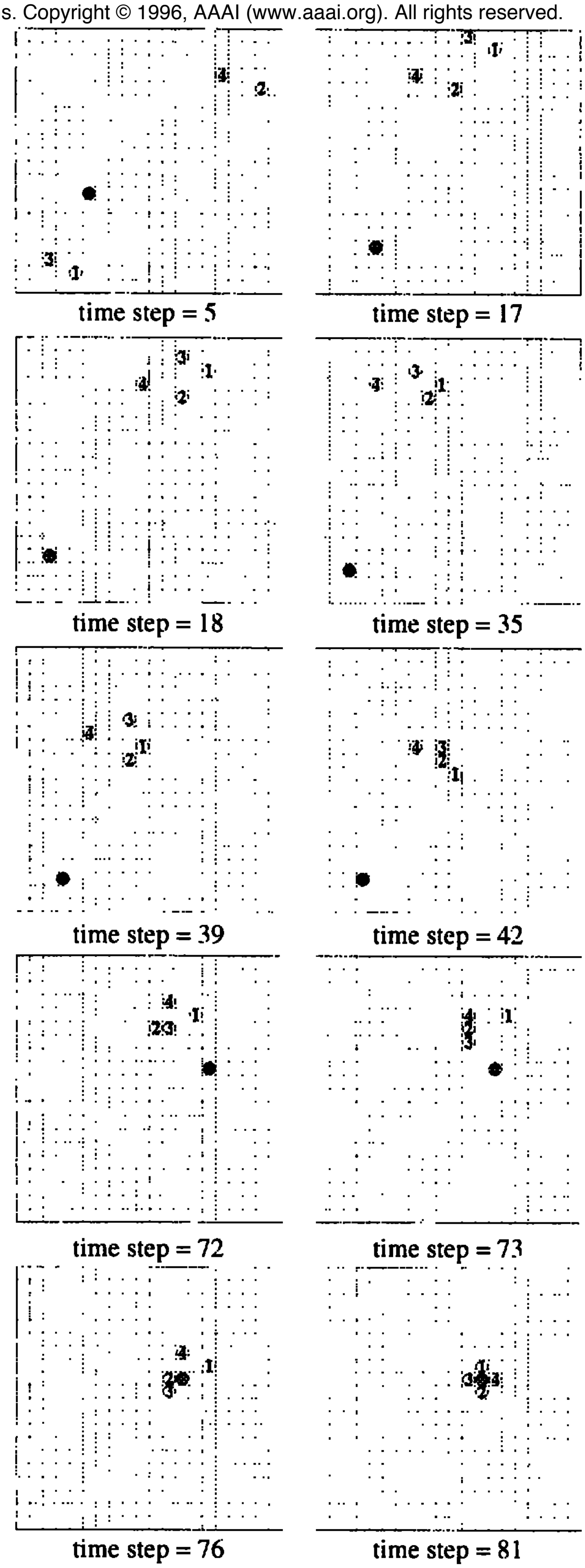
To illustrate this coordinated behavior, a typical prey capturing process by our learning hunters is extracted in Fig. 7. The hunters have already experienced some 10,000 trials in a 20 × 20 toroidal environment. The hunters' visual field depth is 3 and the prey moves randomly.

In this specific trial, at an initial time step of 5, two small groups of hunters are formed, and shortly they are merged into a single large group at time step 18. Then the resulting group move in a consistent direction to collectively search the prey.

At time step 72, the hunter 3 happens to locate the prey within its visual field and starts reaching its own capture position. At this point, the others have not located the prey, but they still attempt to move in a herd. As a result, all the hunters succeed in locating the prey at time step 76, and are attempting to reach their own capture positions effectively. Finally at time step 81, the hunters accomplish their common goal successfully.

So far we have supposed that every hunter has the same visual field uniformly. There is no best one among the 24 possible specializations, and any of them may be developed eventually by the learning hunters. Thus, we can not foresee which specialization will be eventually developed by the learning hunters prior to the simulation run.

What will happen if the hunters have nonuniform



time step = 5

time step = 17

time step = 18

time step = 35

time step = 39

time step = 42

time step = 72

time step = 73

time step = 76

time step = 81

Figure 7: Typical pursuit process by hunters.

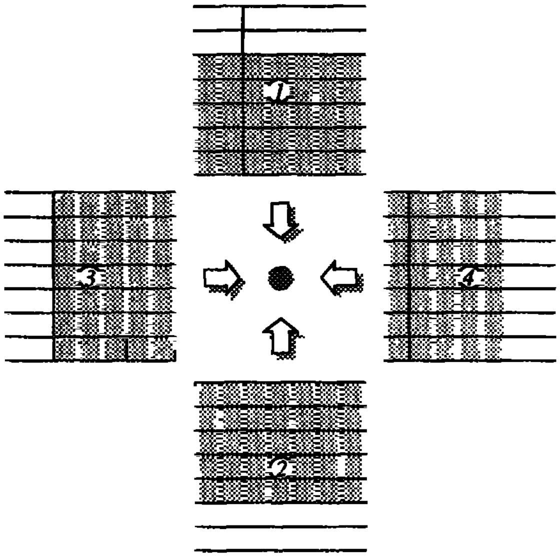


Figure 8: Four hunter agents with nonuniform visual fields. A hunter is indicated by a white circle, whose visual field is depicted by a shaded rectangle surrounding it. At each simulation run, these hunters quickly specialize their functionality and eventually they come to capture the prey, indicated by a black circle, from the direction indicated by their corresponding arrows.

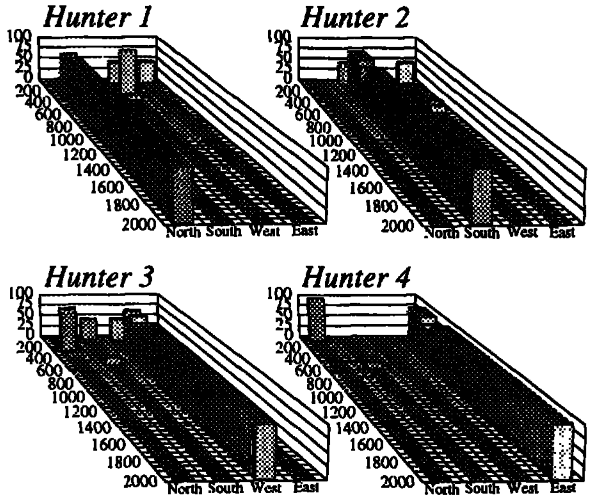


Figure 9: Typical specialization process by hunters with nonuniform visual fields.

visual fields? Are our learning hunters able to steadily identify an appropriate one? We consider one of the most interesting cases where the learning hunters have different visual fields as shown in Fig. 8. In this case, our modular Q-learning hunters developed the same specialization very quickly at every simulation run we performed. A typical specialization process by the hunters is depicted in Fig. 9. Note that they agreed on the eventual specialization much more quickly than the hunters with uniform visual fields.

Synthesis of Coordinated Behavior

Multi-agent reinforcement-learning is a difficult problem, especially when some forms of joint operation are needed for the agents. We do not say our modular reinforcement-learning approach is generally applicable to any multi-agent problem solving. The successful results shown above strongly rely on the attributes of the problem domain and our learning scheme. To suggest this, we explain below how the learning solves our pursuit problem.

Although it might sound strange, our learning hunters acquire the above-mentioned herding behavior first and then specialize their individual functionality. Our hunters learn to make a herd even at initial trials, which depends on some attributes of the problem itself and our learning and state representation schemes.

One of the characteristics of the multi-agent reinforcement-learning problem, like the pursuit problem, is that learning agents can not receive meaningful or positive reward frequently and they are often forced to perform their initial learning without receiving any positive reward, particularly when engaged in some joint tasks. When operating under such a condition, Q-value for each state-action pair experienced by a modular Q-learning agent's learning-modules is repeatedly reduced. Note that the more often the same state-action pair has been experienced, the lower its corresponding Q-value. Thus, at initial trials, the modular Q-learning agents come to not select those actions that take them to frequently-experienced states.

The pursuit problem treated here has similar characteristics. In fact, our hunters have to operate without receiving any positive reward at their initial trials. When its visual field is much smaller than the environment, a hunter often operates without identifying any other agents. Furthermore, any state of its learning-modules where a hunter can not identify any other agents is equally represented by a unique symbol. Note that the larger the environment compared with its visual field, the more often those states represented by the unique symbol are experienced in general. Thus, any state where no other agents are visible comes to seem less desirable for the hunter's learning-modules than the other states. This characteristics of the learning processes encourages the hunters to quite quickly acquire a trait to remain close to other agents, and accordingly get a good chance of capturing the prey. Without acquiring this trait and accordingly herding behavior, our learning hunters can not capture the prey effectively, because they have to occupy all of the four neighbor positions of the prey at the same time.

Concluding Remarks

Recent attempts to let monolithic reinforcement-learning agents synthesize some of coordinated relationships among them scale poorly to more complicated multi-agent learning problems where multiple learning agents play different roles and work together for the accomplishment of their common goals. These learning agents have to receive and respond to various sensory information from their partners as well as that from the physical environment itself. and hence their state spaces grow exponentially in the number of the partners. As an illustrative problem suffering from this kind of combinatorial explosion, we have treated a modified version of the pursuit problem, and shown how successfully a collection of modular Q-learning hunter agents synthesize coordinated decision policies needed to capture a randomly-fleeing prey agent. by specializing their functionality and acquiring herding behavior.

Multi-agent learning is an extremely difficult problem in general, and the results we obtained strongly rely on specific attributes of the problem. But the results are quite encouraging and suggest that our modular reinforcement-learning approach is promising in studying adaptive behavior of multiple autonomous agents. In the next phase of this work we intend to consider the effects of the modular approach on other types of multi-agent reinforcement-learning problems.

References

Benda, M., Jagannathan, V., and Dodhiawalla. R. 1985. On Optimal Cooperation of Knowledge Sources. Technical Report. BCS-G2010-28. Boeing AI Center.

Drogoul, A., Ferber, J., Corbara, B., and Fresneau. D. 1991. A Behavioral Simulation Model for the Study of Emergent Social Structures. *Toward a Practice of Autonomous Systems: Proceedings of the First European Conference on Artificial Life*. The MIT Press.

Gasser, L. et al. 1989. Representing and Using Organizational Knowledge in Distributed AI Systems. In Gasser, L., and Huhns, M.N. eds. 1989. *Distributed Artificial Intelligence*. Vol.II. Morgan Kaufmann Publishers, Inc.

Levy, R., and Rosenschein, J.S. 1991. A Game Theoretic Approach to Distributed Artificial Intelligence, Technical Document, D-91-10, German Research Center on AI.

Ono. N., and Rahmani. A.T. 1993. Self-Organization of Communication in Distributed Learning Classifier Systems. In Proceedings of International Conference on Artificial Neural Nets and Genetic Algorithms.

Ono, N., Ohira. T., and Rahmani, A.T. 1995a. Emergent Organization of Interspecies Communication in Q-learning Artificial Organisms. In Móran. F. et al. eds. 1995. *Advances in Artificial Life: Proceedings of the Third European Conference on Artificial Life*. Springer.

Ono. N., Ohira, T., and Rahmani. A.T. 1995b. Evolution of Intraspecies Communication in Q-Learning Artificial Organisms. In ISCA International Conference: Fouth Golden West Conference on Intelligent Systems. ISCA.

N.Ono. Ikeda. O., and Rahmani. A.T. 1995. Synthesis of Organizational Behavior by Modular Q-learning Agents. In Proceedings of the 1995 IEEE/Nagoya University World Wisepersons Workshop on Fuzzy Logic and Neural Networks/Evolutionary Computation.

Ono. N., Fukumoto. K., and Ikeda. O. 1996. Collective Behavior by Modular Reinforcement-Learning Animats. In Maes. P. eds. 1996. *From Animals to Animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior*. The MIT Press/Bradford Books. Forthcoming.

Rahmani. A.T., and Ono. N. 1993. Co-Evolution of Communication in Artificial Organisms. In Proceedings of the 12th International Workshop on Distributed Artificial Intelligence.

Tan. M. 1993. Multi-agent Reinforcement Learning: Independent vs. Cooperative Agents. In Proceedings of the 10th International Conference on Machine Learning. Morgan Kaufmann.

Yanco. H., and Stein. L.A. 1992. An Adaptive Communication Protocol for Cooperating Mobile Robots. In *From Animals to Animats 2: Proceedings of the Second International Conference on Simulation of Adaptive Behavior*. The MIT Press.

Watkins, C.J.C.H. 1989. Learning With Delayed Rewards. Ph.D.thesis. Cambridge University.

Werner. G.M., and Dyer. M.G. 1994. BioLand: a Massively Parallel Simulation Environment for Evolving Distributed Forms of Intelligent Behavior, In Kitano, H. et al. eds. 1994. *Massively Parallel Artificial Intelligence*. AAAI/MIT Press.

Whitehead. S. et al. 1993. Learning Multiple Goal Behavior via Task Decomposition and Dynamic Policy Merging. In J.H.Connell et al. eds. 1993. *Robot Learning*. Kluwer Academic Press.