Skill Acquisition and Self-Improvement for Environmental Change Adaptation of Mobile Robot

Takashi Minato and Minoru Asada

Dept. of Adaptive Machine Systems
Graduate School of Engineering
Osaka University
Suita, Osaka 565, Japan
minato@er.ams.eng.osaka-u.ac.jp
asada@ams.eng.osaka-u.ac.jp

Abstract

Learning and development are essential processes for an animat to adapt itself to environmental changes so as to accomplish a given task. This paper proposes a single mechanism for learning and self-improvement that results in learning curves similar to the "U-shape" phenomena observed in several psychological experiments concerning the human learning process such as in language acquisition. The basic idea is that (1) the animat monitors its success rate in goal achievement so as to perceive environmental changes instead of relying on signals from a teacher, and (2) in order to reuse acquired knowledge and accelerate reinforcement learning, the animat does not memorize the action values but transfers only the learned policy. The resultant policy (a state transition map where transitions indicate the best actions) may not be optimal in any given environment but it may be able to better handle differences between environments. We apply this model to a mobile robot navigation problem for which the task is to reach the target while avoiding obstacles by means of uninterpreted sonar and visual information. Our experimental results demonstrate the validity of the model.

1. Introduction

Learning and development are essential processes for biological and artificial systems alike. Robots, which may be required to adapt themselves to different environments, provide a typical example. Conventional methods achieve goals in different environments by employing a different module to cope with each kind of environment. However, such methods are limited by the capacity of processing modules. How do biological systems overcome this problem and how could adaptive animats do the same?

In psychology, several experiments indicate the "U-shape" phenomena in the learning of various kinds of skills in humans (Elman et al., 1996). First learning improves monotonically, then performance drops, and finally it rises up again. A typical example can be observed in children who are learning the past tense, that is, to conjugate both regular and irregular verbs correctly (Rumelhart and McClelland, 1986). Researchers had disputed whether a single or a dual mechanism was implicated. An simple solution involves one mechanism for regular verbs and another for irregular. Instead, Rumelhart and McClelland (Rumelhart and McClelland, 1986) showed that a single mechanism plus a carefully designed learning schedule could give the same U-shaped results. Despite much debate and criticism (Pinker and Prince, 1988; Plunkett and Marchman, 1991; Marcus et al., 1992), micro U-shaped curves have been observed in child learning processes for vocabulary development, past tenses of English verbs, physical event cognition, and so on. An artificial neural network has produced similar results (Plunkett et al., 1992), indicating that a single mechanism could cope with different tasks. New tasks were introduced after fixing the policy for the tasks learned so far. That way, the policy that had already been learned was not unlearned while new skills were being acquired.

In this paper, we propose a single mechanism for learning and self-improvement in a mobile robot. The robot must overcome a navigation problem in different environments. The robot learner continually monitors its success rate in achieving the goal in order to perceive changes in the environment when it encounters them. Thus, we distinguish our system from models assumed in several psychological experiments and in artificial neural networks as applied to supervised learning since these models make use of explicit signals from a teacher. The learner does not need to find changes in the environment unless its success rate worsens.

A task domain similar to ours has been dealt with in the lifelong learning area of machine learning (Thrun and Mitchell, 1996). This approach reuses learned policies as *a priori* knowledge to accelerate improvement. Tanaka and Yamamura (Tanaka and Yamamura, 1997) applied a similar idea to a simple grid-world navigational task using a method which combined reinforcement learning with stochastic gradient ascent. We distinguish our method from these because

1. although they accelerate learning, the robot must have a different policy for each kind of environment; our model has only a single policy.
2. more importantly, to learn multiple policies, the robot is explicitly informed of a change in environment; the robot in our model detects a change only when necessary, because the success rate worsens.

To realize a single policy mechanism and accelerate learning, the robot does not keep the action values obtained by reinforcement learning in the previous environment. It only uses the policy (state transition map where transitions indicate the best actions) for action selection.

We also adopt the "learning from easy missions" (LEM) paradigm (Asada et al., 1996) by which the initial positions of the robot are controlled to accelerate the learning. Since LEM is basically considered as a technique for learning in a single environment, we do not deal with it in details here.

The resultant policy obtained by our model does not seem optimal in each individual environment, but may absorb the differences between multiple environments. The remainder of this article is structured as follows. First, the method is explained in details with a brief summary of reinforcement learning, especially Q-learning. Next, the task and some assumptions are given. Finally, we examine the experimental results to test the validity of the model and consider future work.

2. Learning and Self-Improvement

2.1 Basics of Reinforcement Learning

Before getting into the details of our system, we briefly review the basics of Q-learning (Kaelbling, 1993). For a more through treatment, see (Watkins and Dayan, 1992).

We assume that the robot can discriminate the set $\boldsymbol{S}$ of distinct world states, and can take the set $\boldsymbol{A}$ of actions on the world. The world is modeled as a Markov process, making stochastic transitions based on its current state and the action taken by the robot. Let $T(s, a, s')$ be the probability that the system will transit to the next state s' from the current state-action pair (s, a). For each state-action pair (s, a), the *reward* $r(s, a)$ is defined.

Without initial knowledge on T and r, we construct incremental estimates of the action values called Q values on line. Starting with $Q(s, a)$ at any value, usually 0, every time an action is taken update the Q value as follows:

$$Q(s, a) \Leftarrow (1-\alpha)Q(s, a) + \alpha(r(s, a) + \gamma \max_{a' \in A} Q(s', a')). \tag{1}$$

where r is the actual reward value received for taking action a in a situation s, s' is the next state, and α is a learning rate (between 0 and 1).

2.2 Algorithm

As described in **1**, the basic ideas of our method are:

1. by monitoring its success rate, the robot can decide when to restart Q-learning, regardless of the actual change in the environment it encounters (this approach is quite different from existing methods); and,
2. in order to accelerate learning, the action values obtained by Q-learning in the previous environment are not reused for Q-learning in the current environment, but only the policy (action selection) is used. Actually, we have attempted to reuse the action values, but we have often observed that they prevented the robot from learning a new policy.

The algorithm is as follows:

1. Quantize the state space as S.
2. Apply Q-learning to the initial environment, and obtain the policy $P : \boldsymbol{S} \rightarrow \boldsymbol{A}(\boldsymbol{A}$:action set) with the success rate R_s
3. Apply P to any environments unless R_s decrease.
4. If R_s decreases, then find states $\boldsymbol{S}_r \subset \boldsymbol{S}$ where P fails to achieve the goal, and modify P for such states by applying Q-learning as follows until R_s recovers to pre-specified adaptability rate β.
   (a) Apply Q-learning to $\boldsymbol{S}_r \subset \boldsymbol{S}$. Action selection during the learning is as follows:
       if $s \in \boldsymbol{S}_r \cup \boldsymbol{S}_n$ follow the normal action selection in Q-learning, where $\boldsymbol{S}_n \subset \boldsymbol{S}$ denotes inexperienced states.
       else follow P
   (b)
5. Go to 3 with the obtained policy P.

The adaptability rate β determines the extent to which re-learning occurs, that is:

$$R_{sd} = R_{sc} + \beta(R_{sp} - R_{sc}), \tag{2}$$

where R_{sp}, R_{sc}, R_{sd} denotes the success rates in the previous environment, in the current one, and the desired one, respectively.

Based on the selected state vector, we apply the algorithm to the given task with the following specifications:

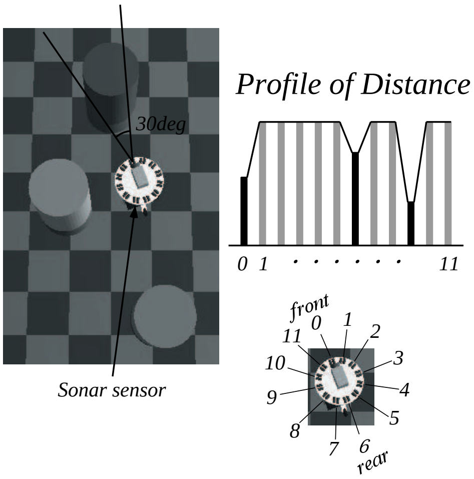
- the learning rate $\alpha = 0.25$, and the discounting factor $\gamma = 0.9$.
- If the robot reaches the target, the positive reward 1 is given. Otherwise 0.
- One trial terminates if the robot reaches the target, makes a collision with any obstacles, or the given time limit expires.

3. Task, Robot, and Assumptions

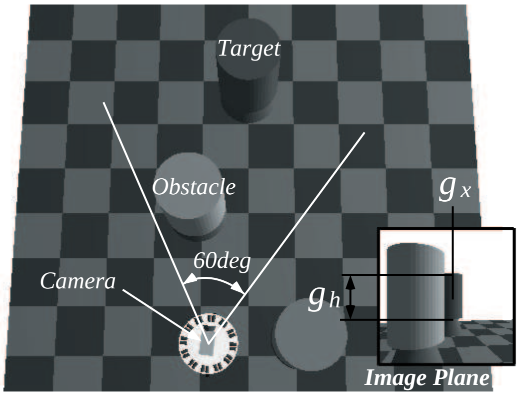
3.1 Our Robot

Our robot has a Power Wheeled Steering (hereafter P-WS) system driven by two motors. We can send commands to each motor independently. In our experiment, we quantized each motor command $\omega_{l(r)}$ into three levels which correspond to forward, stop, and backward, respectively. Totally, the robot has 9 actions.

The robot is equipped with a ring of 12 ultrasonic range sensors (ranging from 0.0 to 300 cm), which have high accuracy for incident angles of less than 15° from the surface normal. The robot is also equipped with a CCD camera. These sensors have their inherent characteristics as follows:



(a) sonar



(b) vision

Figure 1 Sensory information

- **Sonar**
  Using 12 sonar sensors, our robot can sense its surrounding environment in robot centered polar coordinates as a profile of the distance $D_i \quad (i = 0 \sim 11)$ as shown in Figure 1(a). Each sonar sensor in the ring has a field view of roughly 30°. Sonar sensors cannot identify the object (the target or something else).
- **Vision**
  Image processing provides the position and size of the target in the image, even if the object's sensory projections are deformed by occlusion (see Figure 1(b)). However, it is not given information on how to detect obstacles and, therefore, cannot detect them.

3.2 Task

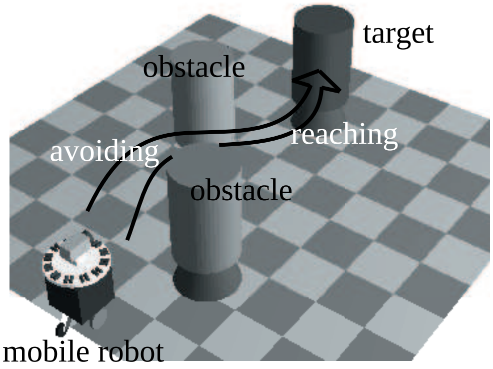


Figure 2 Task and environment

The task of the robot is to reach the target while avoiding obstacles as shown in Figure 2. As mentioned earlier, there are two difficulties with this task:

- visual and sonar information has not been preassigned to specified roles in order to accomplish the given task; therefore, the robot has to learn what kind of information is to be used in which situation. In other words, the sensory data has not been interpreted for the robot, and
- both target reaching and obstacle avoidance tasks have to be achieved simultaneously, through the learning process.

Nakamura et al. (Nakamura et al., 1996) have devised a system that has a limited ability to cope with the above problems in a single isolated environment, but it suffered from the curse of dimensionality: a huge state space. In addition, the environment may change in a few ways here:

1. target and obstacles configuration may change, and
2. the number of obstacles may also change.

Therefore, a learned policy obtained in a single environment may not be applicable to different environments, and usually it takes an enormous amount of time if the robot learns from scratch.

3.3 State Vector Selection

As mentioned above, the state space construction problem is one of the most serious issues in reinforcement learning even in a single isolated environment. Since our robot has a considerably large sensor space, we have to build a reduced-size state space from the original sensor space. As primitive features, we have selected the center position g_x and the height g_h of the target image from vision, and the following from sonar profile (see Figure 3).

- d_{min}: minimum range value
- d_{max}: maximum range value observed
- d_{mean}: mean range value observed
- d_{diff}:$d_{max} - d_{min}$
- θ_{min}: direction of d_{min}
- θ_{max}: direction of d_{max}
- θ_{mean}: mean between θ_{min} and θ_{max}
- θ_{diff}: width between θ_{min} and θ_{max}

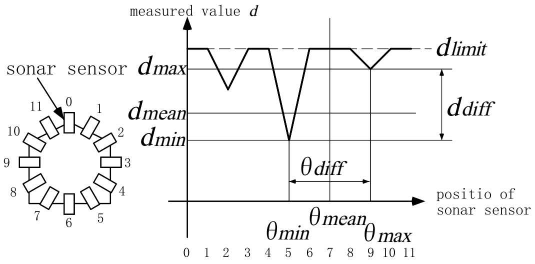


Figure 3 Primitive features from sonar profile

Since these features still constitute a large feature space, we have checked all combinations of state vectors under the constraints of memory space and limited learning time in a single environment, and selected the following state vector $\boldsymbol{x}$ for the task. To focus on the skill acquisition and self-improvement, we skip the details of this procedure, which have been published elsewhere (Minato and Asada, 1998).

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} g_x \\ g_h \\ \theta_{min} \\ d_{min} \end{pmatrix} \tag{3}$$

4. Experimental Results

In order to show the validity of the proposed method, we have applied three kinds of methods to a series of five different environments called A, B, C, D, and E of which top views are shown in Figures 4 and 5 where a solid black circle and gray circles indicate the target and obstacles, respectively. As we can see, the configuration of the goal and obstacles, and/or the number of obstacles are different from each other.

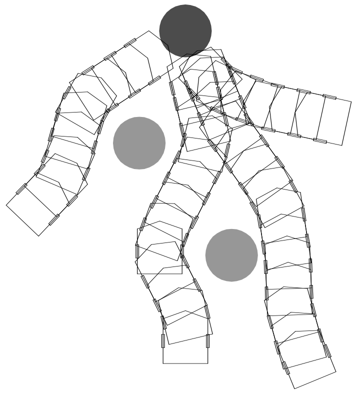
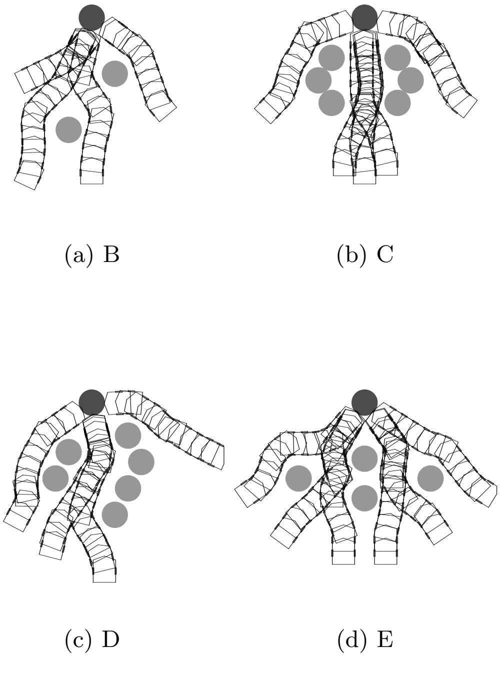


Figure 4 Environment A and successful trajectories



(a) B

(b) C

(c) D

(d) E

Figure 5 Four more environments and successful trajectories

Three methods are:

1. Q-learning with multiple policies: the robot is informed when the environments changes, and all action values are initialized for new learning
2. The proposed method:
3. Another method similar to the proposed one: instead of policy transfer for action selection, all the action values are always retained during the learning.

Each trial terminates when

– the robot reaches the goal,
– the robot makes a collision with any obstacles, or
– the pre-specified period (300 time steps) is expired.

Figure 6 shows the change of the success rate of the Q-learning with multiple policies. The horizontal axis indicates the number of trials. The success rate is measured every 500 trials, the first 200 of which are for normal Q-learning, and the remaining 300 of which are for success rate measuring based on the current policy (action selection is fixed). If the success rate stably achieves the pre-specified one (here 90%), then the policy is fixed and all 500 trials are for success rate measuring. As mentioned in **1**, we applied the LEM (Learning from Easy Missions) paradigm, in this case, three learning stages (easy, moderate, and hard ones) are prepared, therefore two sudden drops can be seen until around the 35,000th trial where the environments changes from A to B, and all the action values are reset to all zeros for new learning in B. Similarly, the robot encounters C, D, and E. Totally, about 145,000 trials are needed for the robot to adapt itself to different environments.

Figure 7 shows the change of the success rate of the proposed method. The pre-specified success rate is 90%, the same as the above, and we set the adaptability rate β as 0.8. The shape of the curve until around the 35,000th trial is completely the same as in Figure 6, but hereafter, the curve has different shape from that by the first method. The changes of the environments have not been informed, but the robot has perceived these changes by monitoring the success rate. Comparing with Figure 6, the total learning time is almost half, and when it encounters A again, the robot has not perceived any changes because the success rate has not dropped down. This implies that the resultant policy is capable of not simply adaptation but also generalization, too.

As the acquired knowledge, the proposed method transfers only the current policy. Further, we may expect to use the action values to accelerate new learning. Figure 8 shows the result of this attempt where it has taken much longer time to adapt itself to B and C, and what's worsen is no convergence can be seen for D until the 200,000th trial. The main reason seems that the state transitions can be different and sometimes opposite, therefore it may take much more time to obtain the correct action values than in case of resetting all action values to zeros.

Examples of the successful trials are shown in Figures 4 and 5 as robot trajectories, some of which do not seem optimal due to a single policy mechanism.

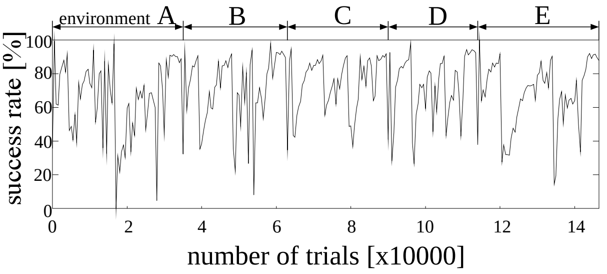


Figure 6 Q-learning with multiple policies

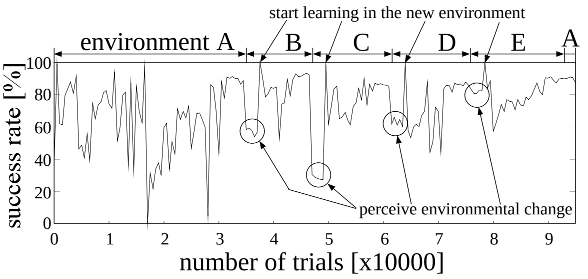


Figure 7 The proposed method

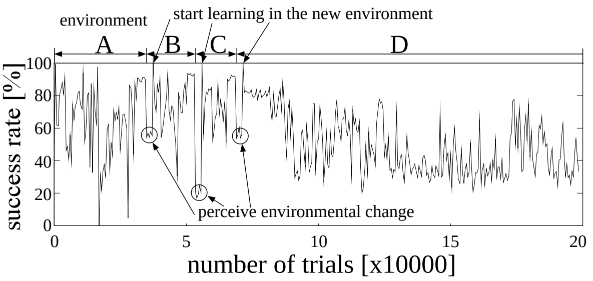


Figure 8 Another method

5. Discussion

We have proposed the model of skill acquisition and self-improvement for a mobile robot to adapt itself to different environments. The learned policy have shown the similar curves in human child learning process often seen in psychological experiments, that is, micro U-shapes. Similar to the method in (Rumelhart and McClelland, 1986), that is, carefully designed input schedule, we have implemented the LEM paradigm to control the order of the situations the robot encounters.

There are many issues to be considered: 1) the definition of the task class in which the robot can gradually skill up the learned policy, 2) state vector selection which is currently off-line process but should be included in online learning process, and 3) real robot experiments.

Acknowledgement

We appreciate Dr. Karl MacDorman, Department of Systems and Human Science, Osaka University for his valuable comments and suggestions.

References

Asada, M., Noda, S., Tawaratumida, S., and Hosoda, K. (1996). Purposive behavior acquisition for a real robot by vision-based reinforcement learning. *Machine Learning*, 23:279–303.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., and Plunkett, K. (1996). *Rethinking Innateness.* The MIT Press, Cambridge, Massachusetts.

Kaelbling, L. P. (1993). "Learning to achieve goals". In *Proc. of IJCAI-93*, pages 1094–1098.

Marcus, G., Ullman, M., Pinker, S., Hollander, M., and T. J. Rosen, F. X. (1992). *Overregularization in language acquisition.* Monographs of the society of for research in Child Development, 57.

Minato, T. and Asada, M. (1998). Environmental change adaptation for mobile robot navigation. In *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems 1998 (IROS98)*, page (submitted).

Nakamura, T., Morimoto, J., and Asada, M. (1996). Direct coupling of multisensor information and actions for mobile robot behavior acquisition. In *Proc. of 1996 IEEE/SICE/RSJ International Conference on Multisensor Fusion and Integration*, pages 139–144.

Pinker, S. and Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28:73–193.

Plunkett, K. and Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perception: Implications for child language acquisition. *Cognition*, 38:43–102.

Plunkett, K., Sinha, C., M∅ller, M. F., and Strandsby, O. (1992). Symbol grounding of the emergence of symbols? vocabulary growth in children and a connectionist net. *Conection Science*, 4(3-4):293–312.

Rumelhart, D. E. and McClelland, J. L. (1986). On leaening the past tenses of english verbs. In Rumelhart, D. E. and McClelland, J. L., editors, *Paralle ditributed processing: Exploration in the micro structure of cognition. Volume 2. Psychological and biological models.* The MIT Press, Cambridge.

Tanaka, F. and Yamamura, M. (1997). An approach to lifelong reinforcement learning through multiple environments. In *6th European Workshop on Learning Robots*, pages 93–99.

Thrun, S. and Mitchell, T. (1996). Lifelong robot learning. Technical Report IAI-TR-93-7, University of Bonn, Dept. of CS III.

Watkins, C. J. C. H. and Dayan, P. (1992). "Technical note: Q-learning". *Machine Learning*, 8:279–292.