Cognitive Systems Research, Vol.3, No.3, pp.459-470, 2002

On the Dynamics of Robot Exploration Learning

Jun Tani

Brain Science Institute, RIKEN

2-1 Hirosawa, Wako-shi, Saitama, 351-0198 Japan

Tel 48-467-6467, Email tani@brain.riken.go.jp

(Cognitive Systems Research, in press)

Original Contribution

Acknowledgment

Requests for reprints should be sent to Jun Tani, Brain Science Institute, RIKEN. 2-1 Hirosawa, Wako-shi, Saitama, 351-0198 Japan

Running Title

Exploration Learning of Robots

On the Dynamics of Robot Exploration Learning

Abstract

In this paper, the processes of exploration and of incremental learning in the robot navigation task are studied using the dynamical systems approach. A neural network model which performs the forward modeling, planning, consolidation learning and novelty rewarding is used for the robot experiments. Our experiments showed that the robot repeated a few variation of travel patterns in the beginning of the exploration, and later the robot explored more diversely in the workspace by combining and mutating the previously experienced patterns. Our analysis indicates that internal confusion due to immature learning plays the role of a catalyst in generating diverse action sequences. It is found that these diverse exploratory travels enable the robot to acquire the adequate modeling of the environment in the end.

1 Introduction

One of the debates in behavior-based robotics is whether or not agents should possess mental processes such as internal modeling, planning and reasoning. Most researchers in behavior-based robotics have rejected the "representation and manipulation" framework since it is widely considered that the representation cannot be grounded and that the mental manipulation of the representation cannot be situated adequately in the behavioral context of the robot in the real world environment. This argument seems to be valid if the agent's mental architecture employs the symbolist framework. One of the major difficulties in the symbolist framework is that the logical inference mechanism utilized in planning or reasoning assumes completely consistent models of the world. This presumption cannot be satisfied if the learning is conducted dynamically in the real world situations. It is, however, also true that the embodiment of mental processes is crucial if we attempt to reconstruct an intelligence at the human level in robots, since even two year-old human infants are said to possess primitive capabilities of modeling and emulating within their adopted environment.

We consider that an alternative to the symbolist framework can be found in the dynamical systems approach (Schoner & Kelso, 1988; Pollack, 1991; Thelen & Smith, 1994; Beer, 1995; Gelder, 1995) in which the attractor self-organized in the phase space plays an essential role in cognitive behavior of the system. Two distinct processes, an internal one and another acting in the environment, organize a unseparable dynamical system when those two are structurally coupled in the phase space. Our previous study (Tani, 1995; Tani, 1996) in navigation learning demonstrated that a robot using a recurrent neural net (RNN) is able to learn the topological structure hidden in the environment, as embedded in its attractor with a fractal structure, from the experiences of sensory-motor interactions. The forward dynamics (Kawato, Furukawa & Suzuki, 1987; Jordan & Rumelhart, 1992) of the RNN generates a mental image of future behavior sequences based on the acquired attractor dynamics. The crucial argument in that study is that the situatedness of the higher cognitive processes are explained on the basis of the entrainment of the internal neural dynamics by the environmental dynamics when those two are coupled by the sensory-motor loop.

A drawback of our study was that the learning was conducted in an off-line manner i.e. the navigation were conducted only after complete learning of the environment. In the current paper, we investigate the developmental processes of acting and learning in the course of the robot's exploration of unknown environment. Our special interest is to investigate how the cognitive processes of the robot could proceed even while its experience and learning are partial and incomplete in the environment. Although there have been number of studies which focus on the problems of exploration learning in robot navigation (Mataric, 1992; Kuipers & Byun, 1993; Yamauchi & Beer, 1996), most of these researches focus on the computational aspects rather than the dynamical system's ones. In the current paper, our model based on the dynamical system approach (Tani, 1995; Tani, 1996) will be further developed by adding the schemes of the consolidation learning and the novelty rewarding. In our experiments using a real mobile robot, although it is still limited in its scaling, interactive processes among rehearsing, consolidation, on-line planning, acting and rewarding will be closely observed. Our analysis on the experimental results will show some essential characteristics of the exploratory learning as articulated using the dynamical systems language.

2 The Model

In this section we introduce a neural net model which enables the system to perform exploratory behavior, goal-directed planning and behavior-based learning. The

neural net architecture employed has been built by combining pre-existing neural net schemes. In the learning process, both reinforcement learning and prediction learning are conducted (Werbos, 1990). Using reinforcement learning, the action-policies for better rewarding are reinforced, through which the most preferred action in the current state is selected. In prediction learning, the forward model (Kawato, Furukawa & Suzuki, 1987; Jordan & Rumelhart, 1992) is adapted to extract the causality between the action and the sensation. In goal-directed planning, the inverse dynamics scheme (Werbos, 1990; Jordan & Rumelhart, 1992) is applied to the forward model in order to generate possible action sequences. In this planning process, the action preferences adapted using reinforcement learning provides heuristics for searching for the better rewarded action sequences. In the current formulation, rewards are given to the system based on the novelty which the system experiences for each exploration action (Schmidhuber, 1991; Thrun & Moller, 1992). In other words, when the system cannot predict the next sensation in terms of the current action, the current action is rewarded. In addition, the prediction learning attempts to learn to predict how much prediction error it will make (Schmidhuber, 1991; Thrun & Moller, 1992). By combining this novelty-rewarding scheme with the reinforcement learning and with the prediction learning schemes, the system tends to explore the workspace regions with which it is unfamiliar. Through the consolidation learning and rehearsing (Tani, 1998), certain inconsistency could remain in generating the internal model of the environment since the novelty rewarding scheme continues to bring new experiences to the system which might occasionally conflict with the previous experiences of the system. As the results, the action selections might not be optimized as always because of the incompleteness in the acquired internal modeling. The main purpose of this modeling is to investigate the possible interplay between exploration and learning when the system develops based on this sort of the unstationary dynamics.

Fig 1 shows schematically how multiple cognitive processes interact in our proposed model. Action sequence is generated by means of the on-line planning with the novelty rewarding scheme during the exploratory travel. The episode in terms of the sequence of sensory-action pair is once stored in the short-term memory. After the termination of one exploratory travel, the episodic sequence stored in the short-term memory is consolidated into the long-term memory while the memory rehearsal takes place. The modification of the long-term memory affects the way of planning and action generation in the later exploratory travels.

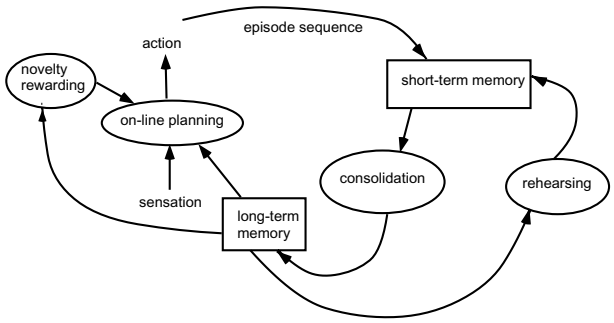


Figure 1: Interactions among multiple cognitive processes.

2.1 The neural net architecture

A RNN architecture is employed in our model as shown in Fig 2. The RNN is operated in a discrete time manner with synchronizing with each event (the events in our navigation task correspond to occasions of encountering a corner, as will be described in a later section.) At the tth event step, the RNN receives the current sensory input s_t, the current reward signal r_t, and the current action x_t. The RNN then outputs the prediction of the next-event sensory input $\hat{s}_{t+1}$, the reward signal $\hat{r}_{t+1}$, and its preference for the next action $\hat{x}_{t+1}$ which is expected to obtain the maximum reward in the future. In the novelty rewarding scheme, the normalized square of the current prediction error for the sensory inputs is used to evaluate the current novelty reward. It is noted that the reward is generated internally from the prediction error of the RNN and the RNN itself is forced to learn to predict it. The RNN has context units c_t in the input and output layers in order to account for the internal memory state (See Ref. (Tani, 1996) for more details of the role of context activation in navigation learning.)

(A) Learning: The RNN learns to predict the next sensory inputs and the rewards corresponding to the current sensory inputs and the action selection. By this means the internal model of the environment is acquired in terms of the forward model. The preference for the next action is learned by a variant of the profit sharing method (Holland & Reitman, 1978) by which sequences of actions which lead to unpredictable experiences are reinforced. Both learning processes are executed in the RNN using the back-propagation through time (BPTT) algorithm after each exploratory travel is terminated. During each travel, the actual sequence of the sensory inputs, the novelty reward, and the action outputs are stored in the short-term memory, which are used for the later consolidation learning processes. For the reinforcement learning, the

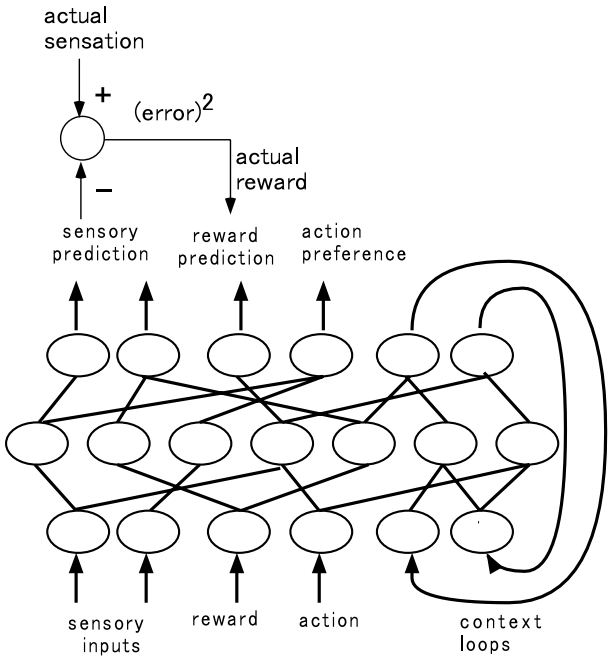


Figure 2: The RNN architecture.

contribution factor of action selection at each step for the future cumulative reward is computed using the sequence of reward experienced during the travel. The contribution factor at lth step: cf_l is computed as:

$$cf_l = \sum_{i=l}^{\tau} \alpha^{i-l} \cdot r_i \tag{1}$$

where τ represents the terminal step of the travel and α is the decay coefficient of the reward in time. The sequence of actions are reinforced as proportional to the obtained contribution factor. In the BPTT learning scheme, the action sequence which has been taken in the previous travel is used as target outputs of actions. The learning error of the action outputs at lth step is multiplied by the contribution factor cf_l which is back-propagated through time. This means that actions which have contributed higher for receiving the novelty reward later are learned with the higher pressure.

(B) Incremental learning by consolidation: It is difficult for RNNs to learn incrementally the sequences given. It is generally observed that the contents of the current memory are severely damaged if the RNN attempts to learn a new teaching sequence. Therefore, the previously described schemes of the forward model learning and reinforcement learning are combined with the framework of the consolidation learning. Observations in biology show that some animals and humans may use the hippocam-

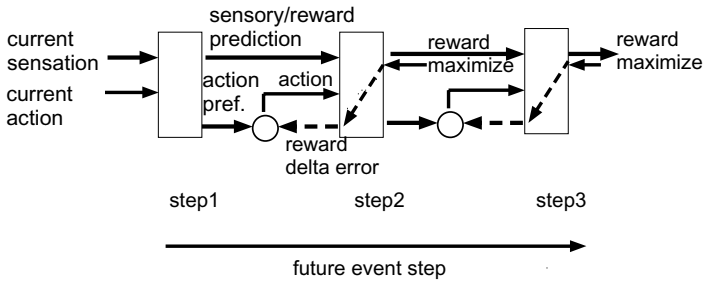


Figure 3: Plan of action sequence is dynamically generated by cascading the RNN for future event steps. Action at each step is determined utilizing both contributions from the action preference and the delta error for maximizing the predicted reward.

pus for temporary storage of episodic memories (Squire, Cohen & Nadel, 1984). Some theories of memory consolidation postulate that the episodic memories stored in the hippocampus are transferred into some regions of the neocortical systems during sleep. McClelland, McNaughton and O'Reilly (1994) further assume that the hippocampus is involved in the reinstatement of the neocortical patterns in long term memory and that the hippocampus plays a teaching role in training the neocortical systems.

We have applied these hypotheses to our model of RNN learning with considering that the RNN stores the long term memory (Tani, 1998). In our system, a new episodic sequence experienced in the current travel is once stored in the "hippocampal" database. (This is implemented not by a particular neural network modeling but by a simple programming.) In the consolidation process after each travel, the RNN generates the imaginary sensory action sequence by rehearsing based on its previous learning. This rehearsal can be performed by repeating "planning" without actual executions of the generated action sequences. The forward dynamics of the RNN with the sensory closed-loop (i.e. the sensory prediction outputs fed-back to the sensory inputs) can generate the imaginary sequence (as will be described in the next section). Then, the RNN is re-trained using both the new episodic sequence stored in the "hippocampal" database and the rehearsed sequences generated from the RNN simultaneously using the learning scheme previously described. This combination of rehearsal and learning allows the memory system to be re-organized without suffering from some catastrophic interference between the novel experiences and the pre-learned memory.

(C) Planning: The objective of planning is to find the action plan $x* = (x_0, x_1, ...x_\tau)$ which generates the path to maximize the future cumulative discounted rewards. Fig 3 shows the scheme of the planning. The RNN is transformed into a

cascaded feed-forward network consisting of τ_{max} steps. The action sequence is dynamically computed by using contributions both from the forward model part and from the action preferences part. Inverse dynamics (Jordan & Rumelhart, 1992) are applied to the cascaded RNN in order to obtain the update of the action plan $\triangle x*$ for maximizing the cumulative discounted reward expected in the future sequence. We consider the following energy function by taking the negative of the expected cumulative discounted reward from the current time step to the terminal step τ:

$$Em(x*) = -\sum_{i=0}^{\tau} \alpha^i \hat{r}_{i+1} \tag{2}$$

where α is the decay coefficient of the future expected reward. τ represents the terminal step where the RNN predicts the situation that the robot goes out of the workspace boundary. This means that in (2) the expected reward at each step is summed up as long as the robot is expected to be within the boundary of the workspace. The exact implementation of this out-of-boundary situation in our robot experiment will be explained in the later section.

The back-propagation through time (BPTT) algorithm (Rumelhart, Hinton & Williams, 1986) is used to compute the update to the action sequence which minimizes the energy assumed in the model part. Firstly the forward computation is conducted on the cascaded network, in which the lookahead prediction of τ_{max} steps for the temporal action program $(x_0 x_1 x_2 \cdots x_{\tau_{max}})$ is obtained. In this lookahead prediction, if the out of the boundary situation is predicted at step τ, the cascade in the forward computation is terminated at this step. Next, an update of the action at each step is obtained. The gradient of Em in (2) with respect to each action $x_n (0 \leq n \leq \tau - 1)$ is calculated; this indicates the direction of update for the action. The update is obtained by means of back-propagating the error between the desired reward and the predicted reward to action nodes in the cascaded network. Here, the error between the desired reward and the predicted reward is obtained as $1.0 - \hat{r}$ for each step. This back-propagation proceeds through the cascaded network from step τ to step 0.

In addition to the contribution from the model part, the action preference influences the planning dynamics in that the difference between the preferred action and the planned action at each step is minimized. The update to the action at each future step is obtained by taking the sum of both parts of the contributions and adding a Gaussian noise η. The update to the action plan is therefore

$$\triangle x_i = \epsilon \cdot [\frac{-\delta Em(x*)}{\delta x_i} + kr \cdot (\hat{x_i} - x_i) + kn \cdot \eta] \tag{3}$$

Here, the first term represents the contribution from the forward model prediction of the reward. The second term represents the contribution from the action preferences. The Gaussian noise term is employed in the third term to prevent the plan dynamics being captured in a local minimum. The value of kn is changed in linear proportion to the value of Em. Therefore the plan search dynamics become stabilized when the energy is minimized; otherwise, it continues to be activated. Here, the reader is reminded that the contributions to the update from the forward model and from the action preferences do not always agree with each other in the course of the exploration processes since the overall system dynamics are characterized by highly nonlinear and non equilibrium dynamics.

Our emphasis in the presented scheme is that the planning process proceeds totally in an autonomous and dynamic manner while the system interacts with the environment. The plan updated by (3) is continuously computed in a real time manner while the robot explores the workspace. A once generated action plan, which is settled in an energy minimum, could be dynamically re-organized when the prediction in the plan does not agree with the real sensation. In this manner, the planning processes can be re-situated autonomously in the behavioral interaction with the environment.

3 Experiment

3.1 Task setting

A mobile robot as shown in Fig 4 is used for the experiment. The robot is equipped with an infrared type range sensor belt on its body. As a default behavior, the robot continues to travel the workspace by wall-following. When the robot encounters a corner, it determines whether it will continue to follow the current wall on its left side or instead to leave the current wall after turning the corner and to move forward diagonally at 45 degrees to the right until it encounters another wall. In this setting, the action can be represented by one bit of information which represents whether or not to branch at the branching points (corners). The RNN architecture receives the travel vector as its sensory inputs at each branch point. The travel vector represents what distance and from which direction the robot has traveled since the previous branch. These values are measured by taking the sum and the difference between the left and right wheel's rotation angles.

Fig 5 shows the adopted workspace for the experiment. The robot starts its exploration travel from a fixed home position and the exploration is terminated when the

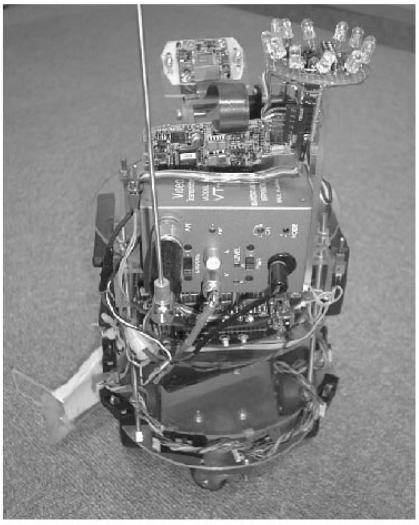
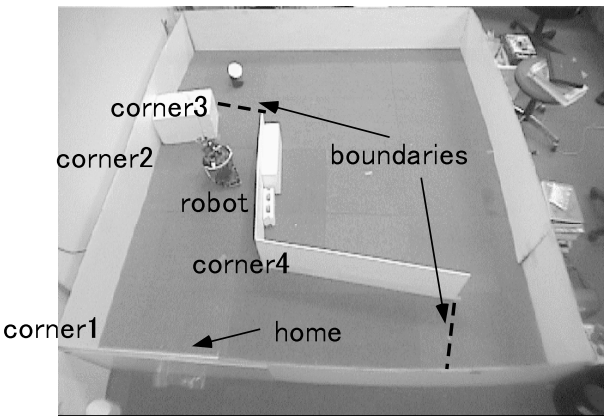


Figure 4: The robot employed in the experiment.



(b)

Figure 5: The adopted workspace. Dotted lines show the predefined boundary of the workspace.

travel takes it outside a predefined boundary. (The home position and the boundary predefined in our experiment is shown in Fig 5.) At the moment of termination, the RNN receives the termination signal in its sensory input and the robot is brought back to the home position manually. (In the planning process, if the RNN predicts the sensation of the terminal signal at a certain lookahead step τ, the plan is inhibited to go forward beyond the step.) Following this, the consolidation process takes place in which the RNN repeated rehearsing of 10 imaginary travels. After the consolidation, exploration by the robot is resumed.

3.2 Results

The robot repeated the exploration travels 20 times in the experiment. This experiment was conducted three times under the same conditions. Fig 6 represents how the average prediction error changes as the exploration travel is repeated in the three experimental cases. Although, there are certain oscillations in the time course of the prediction error, the error after the 16th travel seems to be minimized in all three cases. It can be said that the prediction error converges after enough repetitions of the travel. (There could be certain sudden rises of the error even after the convergence, as is observed in the 15th travel in the experiment-2. It is considered that such rises of the error after the convergence are due to noise accompanied with the real world experiments in most cases.)

In the following, we examine how the diverse travel sequences are generated in the course of exploration. Fig 7 shows all 20 trajectories of the robot's travel observed in one experimental case (experiment-1). In the initial period of the exploration, the robot tends to repeat the same branching sequences. As is evident in Fig 7, the same trajectory is repeated for the first two travel sequences. For the third sequence, branching changes and a different trajectory is generated. This trajectory is repeated in the next two travel sequences. The trajectory in the sixth travel sequence seems to be generated by combining the two travel sequences previously experienced. We summarize that the novelty rewarding scheme causes the observed repetitions and variations in the travel. When the robot undergoes a previously unexperienced travel sequence, the branching sequence experienced is reinforced strongly because of its unpredictability. When the same trajectory is repeatedly generated through reinforcement, the sequence becomes predictable and is rewarded less. As a result, the probability of modifying the current travel is increased.

It is interesting to observe the rehearsing during the consolidation learning since

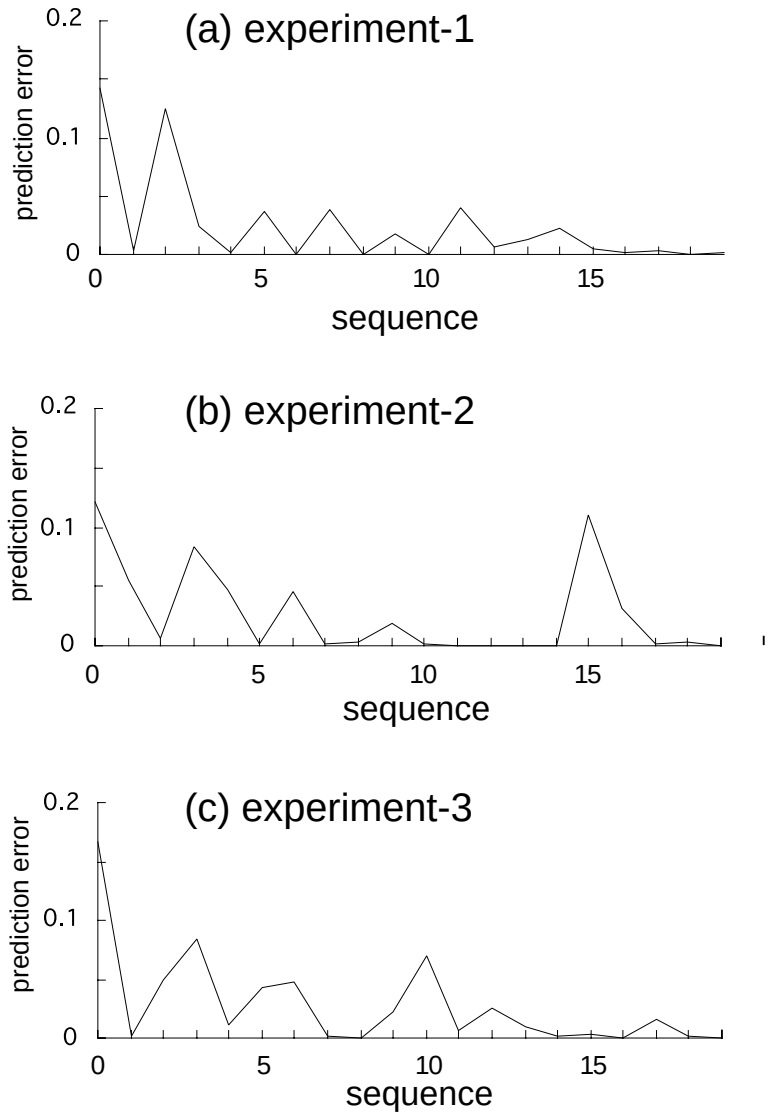


Figure 6: The time history of the average prediction error through the sequence of repeated exploratory travels. It is shown for the three experiment cases with the same condition.

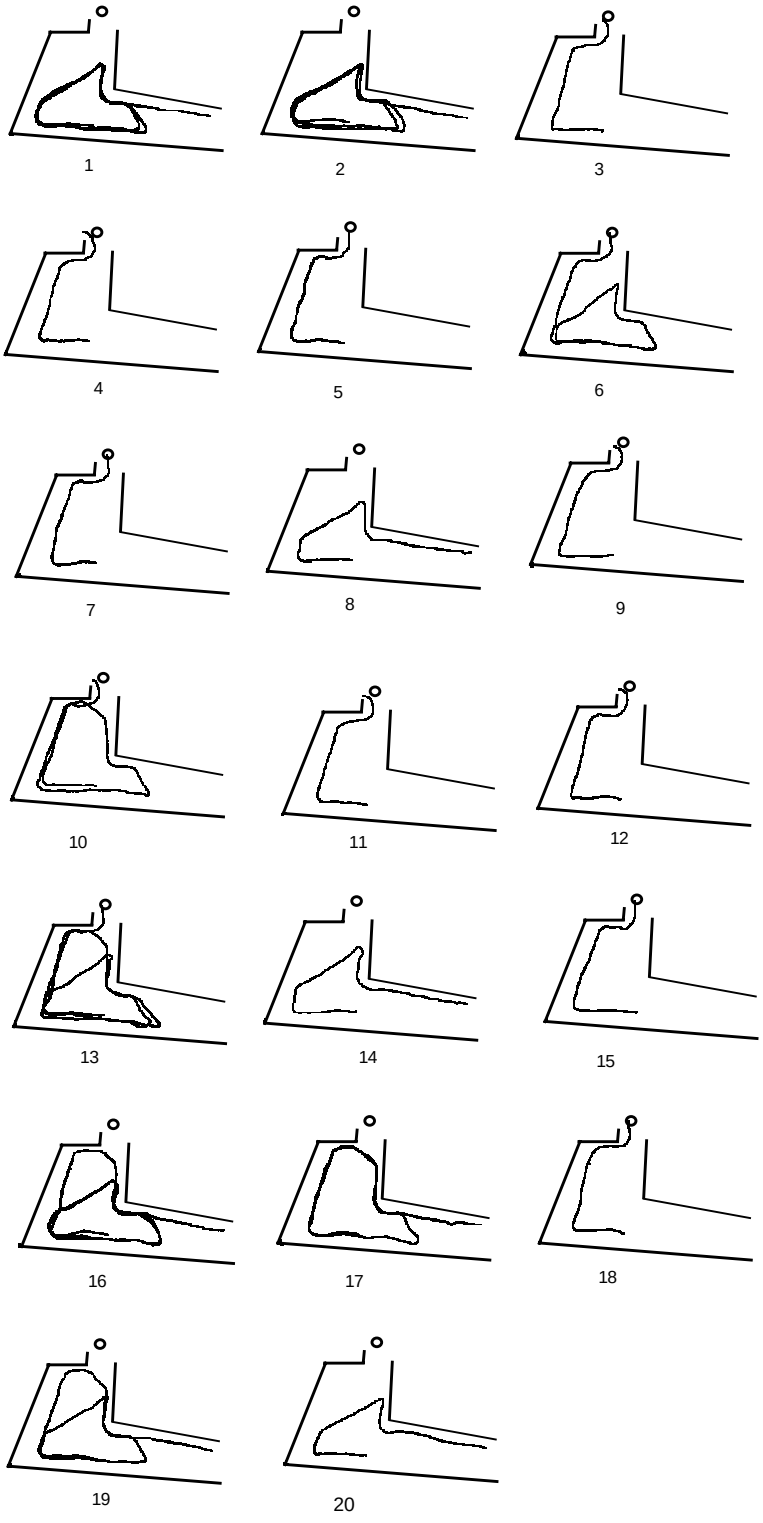


Figure 7: The trajectories of the robot exploration travel for one experimental case. The travel sequence number is given.

the contents of the rehearsing activities represent what the robot has learned so far. Fig 8 shows how the diversity of the rehearsed plans at each consolidation learning process change as the exploration proceeds. The upper graph in the figure represents the corresponding predicted rewards of the plans generated; the lower graph shows ID of all rehearsed plans generated during each consolidation learning period. (The ID is assigned for each plan generated by encoding the bit pattern of the branching sequence, a maximum of 10 time steps in length, into numbers from 0 to 512.) It is observed that the diversity of plans is increased and that the predicted reward is decreased as the exploration trial is continued. We observed that the rehearsed plans are generated not just by repeating the sequences previously experienced but by combining previously experienced sequences into new ones. This way of increasing the diversity in the memory rehearsing quite agrees with that of the travel trajectories as shown previously.

An interesting question is how novel action sequences are generated in the planning process. What we found is that novel branching sequences are originated not merely by the noise term in the planning dynamics but also by the internal confusion caused by the incremental learning. This point is illustrated by considering an example seen in the 10th travel sequence in the experiment-1. In this travel sequence, the robot, starting from the home position, continued to follow the wall after passing corner1, then it branched to another wall after passing corner2 (See Fig 9.) This branching at corner2 is a novel experience for the robot. We investigated how this branching decision was generated by examining the recorded planning process. Fig 10 shows the actual planning processes which took place immediately before the branching was made at corner2. In Fig 10 (a) each column consisting of white and black squares represents a branching sequence plan at each time step of the planning process, where the black and white squares denote branching and non-branching, respectively. Fig 10 (b) indicates the predicted reward for the plan generated. At the beginning of the planning process, a plan of not branching twice is generated with a low predicted reward. This plan will repeat the 5th travel sequence if actually realized. At the end of the planning process, plans are generated such that branching actions are planned to occur repeatedly after passing corner2 with an expectation of a higher reward, even though such action sequences have never been experienced. It is noted that this type of plan was not observed when the robot approached the same corner in its earlier travels. Further examination showed that the lookahead prediction of the sensory sequences after branching at corner2 and at corner1 are mostly the same. This can be interpreted as meaning that the robot hypothesized that branching at any corner would lead to

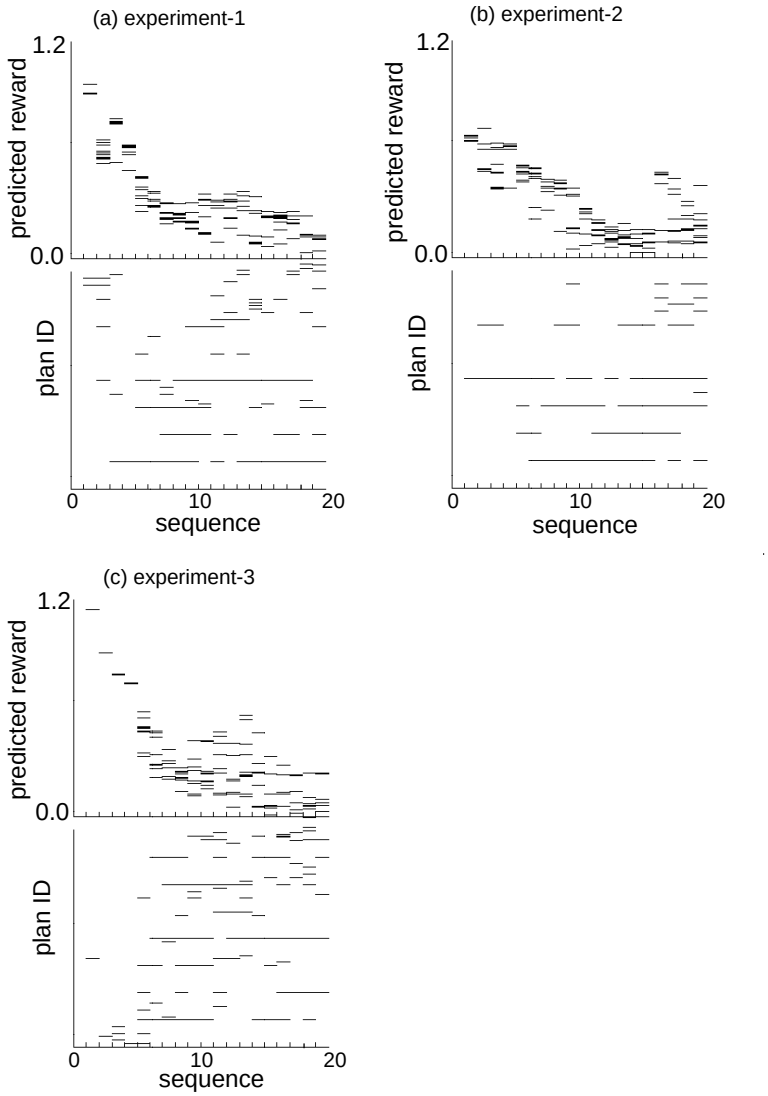


Figure 8: Changes in the diversity of the rehearsed plans during the three exploration experiments.

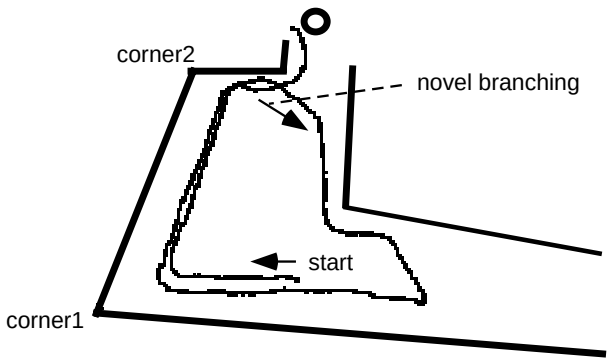


Figure 9: In the 10th travel, the robot made a novel branching after passing corner2.

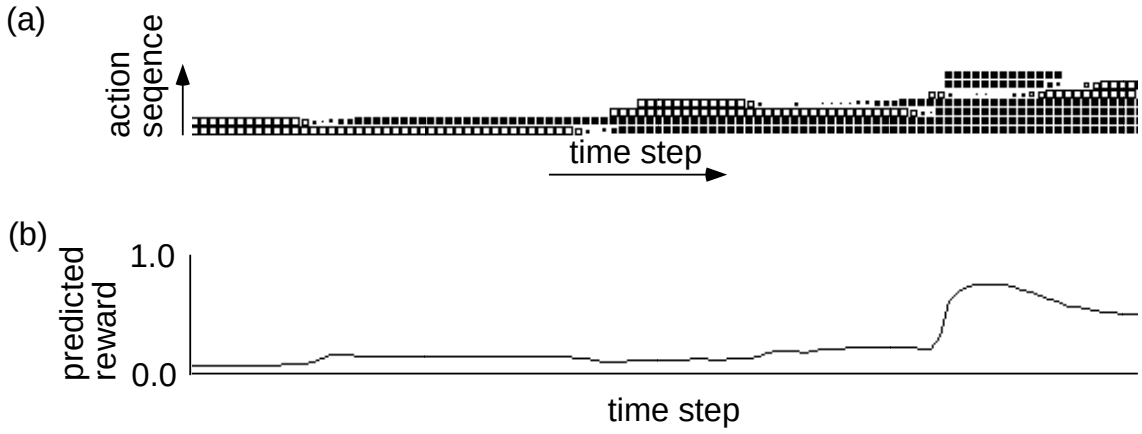


Figure 10: (a) The time history of plan generation at the corner2, (b) the predicted reward of the corresponding plan.

better chances for encountering novel experiences because it applied the situation after branching at corner1 to consider the situation at corner2. (Indeed, the travel will continue as long as branching is selected at approaching corners without terminating the travel by going out of the workspace boundary.) We conclude that the novel action of branching at corner2 results from the expectation of a higher reward which is falsely anticipated by means of fake memory generated in the course of consolidating immature experience. This phenomenon of the novel action trial being generated by fake memory and the internal confusion was seen frequently in the middle of the learning process.

Finally, we investigated how the internal modeling develops by examining the evolution of the RNN attractor. The phase plots were drawn by iteratively activating the RNN in the closed-loop mode with inputs comprising 4000 steps of arbitrary branching action sequences generated under the constraints that the robot does not go beyond the boundary of the workspace. In order to project the context units activation into the two dimensional space, c_1 and c_2 values are taken for average activation over one half of the context neurons and that over the other half of the context neurons, respectively. Fig 11 shows the attractor which appeared in the phase plots generated at different stages in experiment-1. In Fig 11, cluster structures consisting of multiple segments are clearly seen in the later periods of the exploration travel. Our examination clarified that this set of cluster segments represents the global attractor. Further analysis indicated that in the phase plots in Fig 11 (c) and Fig 11 (d) there are correspondences between the segments and the branching position in the workspace and also that the graph structures are topologically equivalent between that of the state transition in the phase space and that of branching of the robot trajectories in the environment. The trajectory of the state transitions in the RNN forward dynamics is closed in the phase space while the toplogical trajectory of the robot navigation is closed in the real environment, as we have shown previousely (Tani, 1996). In this condition, it is said that the "dynamical closure" is generated in the RNN dynamics (Tani, 1996). Here, the dynamical closure emerges as the global attractor by which the forward dynamics of the RNN can generate stable and sound sensory predictions for the possible action sequences even in non-Markovian environment.

However, such structures were barely seen in the phase plots in Fig 11 (a) and Fig 11 (b). While the learning process is "immature", the shape of the attractor varies substantially after each learning and neural dynamics exhibits diverse trajectories in the phase space and the robot behaves as if it were confused. In the meanwhile, the attractor develops step by step as the diverse exploration repeated and finally the dynamical closure is organized in the internal neural dynamics. This phenomena can be

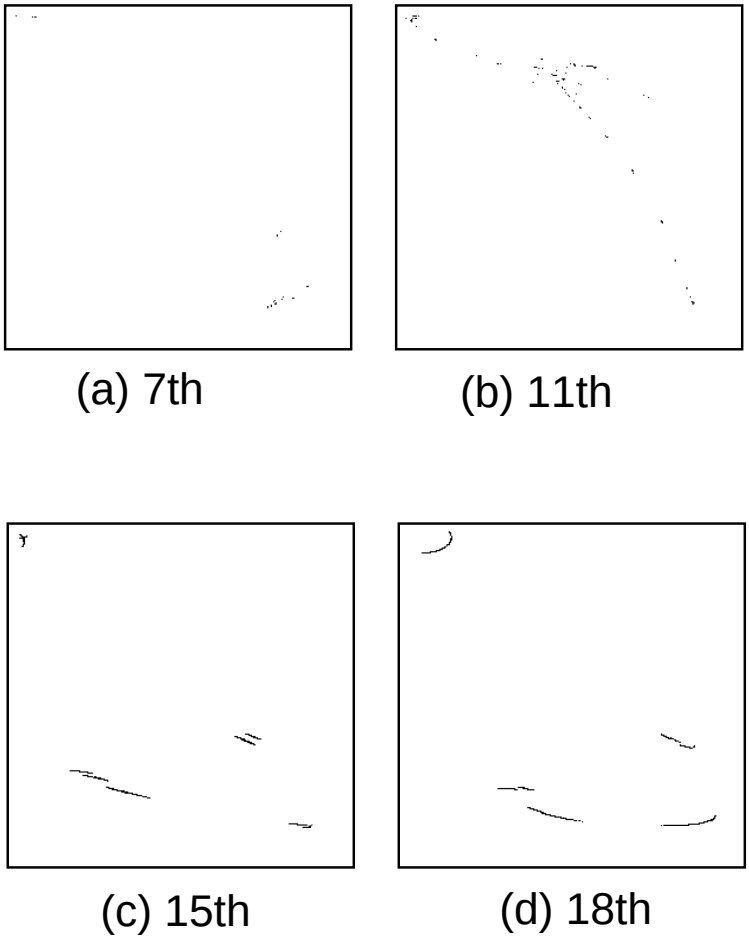


Figure 11: The RNN attractor appeared at a certain stage of the learning process in experiment-1. The learning stage is given at the base of each plot.

partially explained by the prior works by Pollack (1991) and Taiji & Ikegami (1999) in RNN learning of finite state machines (FSM). They have shown that the target (FSM) is adequately embedded in RNNs when its attractor appears to be in a segments-like low dimensional shape while the complexity in the sequence generation of the RNN tends to be much higher than that of the target FSM when the attractor to be in a cloud-like shape as shown in Fig 11 (b).

4 Conclusion

In the experiments, it was shown that the robot learned incrementally about its workspace through exploration and that the robot was eventually successful in obtaining a sound model of the workspace. However, the emphasis in our study is on the observation of dynamic processes before the sound model is achieved. In the beginning, a few travel sequences are repeated and later some combinations of them are made, which were observed both in the memory rehearsing processes and in the actual robot travels. In the middle period, novel actions are frequently tried with a false expectation of the future consequences. The confusion due to the immaturity turns out to be beneficial since it acts as a catalyst for generating the diverse behavior required to explore the environment. The repeated learning of such diverse behaviors enables the robot to acquire the sound model later.

Our study has shown a novel view of re-interpreting the "representation and manipulation" framework. The conventional idea was that a complete internal model exists and then a consistent mental operation using the model is guaranteed. In our idea, the internal model never exists in terms of static entity, but it appears as a dynamical entity in the mental processes of memory rehearsing and planning. It is also fair to say that most of the dynamical systems approaches for cognition tends to focus on the system's direction moving towards states of coherence and convergence. The current study as well as the previous study (Tani, 1998) have indicated that the direction towards states of incoherence and diversity is equally essential. Our crucial argument is that the ultimate autonomy of cognitive systems would take place in such fluctuated processes as the results of highly nonlinear and non-equilibrium interactions between the mental processes and the physical processes of acting.

Finally, the scalability of the current scheme is discussed. It is fair to say that our experiments were conducted under a simple setting of the environment. Our preliminary experiments with changing the complexity of the workspace showed that the exploratory incremental learning hardly converged when the number of branching points

exceeds 7. The attractor shape continued to change and the averaged prediction error cannot be minimized even after substantial number of exploratory travels were repeated. However, in such situations, the robot can travel with good prediction as long as the robot keeps track of familiar trajectories. It is assumed that the predictability could be maintained while combinations of familiar trajectories are repeated, for example, by cutting off the novelty seeking behavior even the complexity of the environment is increased. Further detailed studies are expected for seeking (1) the possible relations between the degradation in the rationality of behavior and the complexity of the environment, (2) the cognitive mechanism which enables the limited rational behavior utilizing incomplete modeling of the environment.

References

[1] Beer, R. (1995). A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence*, 72(1), 173–215.

[2] Holland, J.H. & Reitman, J.S. (1978). Cognitive systems based on adaptive algorithms. In: D.A. Watermann and F. Hayes-Roth (Eds.), *Pattern Directed Inference Systems*. New York: Academic Press.

[3] Jordan, M.I. & Rumelhart, D.E. (1992) Forward models: supervised learning with a distal teacher. *Cognitive Science*, 16, 307–354.

[4] Kawato, M., Furukawa K. & Suzuki, R. (1987) A hierarchical neural network model for the control and learning of voluntary movement. *Biological Cybernetics*, 57, 169–185.

[5] Kuipers, B. & Byun, Y. (1993) A robot exploration and mapping strategy based on a semantic hierarchy of spatial representation. *J. Robot Autonomous Sys.*, 8, 47–63.

[6] Mataric, M. (1992). Integration of representation into goal-driven behavior-based robot. *IEEE Trans. Robotics and Automation*, 8(3), 304–312.

[7] McClelland, J.L., McNaughton, B.L. & O'Reilly, R. (1994). Why there are complementary learning systems in the Hippocampus and Neocortex. Technical Report PDO.CNS.94.1, Carnegie Mellon University.

[8] Pollack, J. (1991). The induction of dynamical recognizers. *Machine Learning*, 7, 227–252.

[9] Rumelhart, D., Hinton, G., & Williams, R. (1986). Learning internal representations by error propagation. In: D.E. Rumelhart and J.L. McClelland (Eds.), *Parallel Distributed Processing*. Cambridge, MA: MIT Press.

[10] Schmidhuber, J. (1991). A possibility for implementing curiosity and boredom in model-building neural controllers. In: J.A. Meyer and S.W. Wilson (Eds.), *From Animals to Animats: Proc. of the First International Conference on Simulation of Adaptive Behavior*, pp. 222–227. Cambridge, MA: MIT press.

[11] Schoner, G & Kelso, J.A. (1988). *Dynamic pattern generation in behavior and neural systems. Science*, 239, 1513–1520.

[12] Taiji, M. & Ikegami, T. (1999) Dynamics of internal models in game players. *Physica D*, 134, 253–266.

[13] Smith, L., & Thelen, E. (1994). *A dynamic systems approach to the development of cognition and action.* Cambridge, MA: MIT Press.

[14] Squire, L.R., Cohen, N.J. & Nadel. L. (1984) The medial temporal region and memory consolidation: A new hypothesis. In: H. Weingartner and E. Parker, editors, *Memory consolidation*, pp. 185–210, Hillsdale, N.J: Erlbaum.

[15] Tani, J. (1995) Essential dynamical structure in a learnable autonomous robot. In: *Proc. of the Third European Conf. of Artificial Life (ECAL'95).*

[16] Tani, J. (1996). Model-Based Learning for Mobile Robot Navigation from the Dynamical Systems Perspective. *IEEE Trans. System, Man and Cybernetics Part B*, 26(3), 421–436.

[17] Tani, J. (1998). An interpretation of the 'self' from the dynamical systems perspective: a constructivist approach. *Journal of Consciousness Studies*, 5(5-6), 516–42.

[18] Thrun, S.B. & Moller, K. (1990) Active exploration in dynamic environments. In *in Proc. of NIPS 4*, 531–538.

[19] van Gelder, T. & Port, R. (1995) It's about time: an overview of the dynamical approach to cognition. In R. Port and T. van Gelder (Eds.), *Mind as Motion*, pp. 1–44. MIT Press, Cambridge, MA.

[20] Werbos, P. (1990) A menu of designs for reinforcement learning over time. In W.T. Miller, R.S. Sutton, and P.J. Werbos (Eds.), *Neural Networks for Control*, pp. 67–95. MIT Press, Boston, MA.

[21] Yamauchi, B.M. and Beer, R.D. (1996) Spatial learning for navigation in dynamic environment. *IEEE Trans. Syst. Man Cybern.*, 26(3).