Appeared in Journal of Consciousness Studies, Vol.5, No.5/6, pp.516-542, 1998

An Interpretation of the "Self" From the Dynamical Systems Perspective: A Constructivist Approach

Jun Tani *
Sony Computer Science Laboratory Inc.
3-14-13 Higashi-gotanda, Tokyo, 141 JAPAN. tani@csl.sony.co.jp
Sony CSL Technical Report: SCSL-TR-98-18
To appear in Journal of Consciousness Studies, 5(5-6), 1998.

October 6, 2011

Abstract

This study attempts to describe the notion of the "self" using dynamical systems language based on the results of our robot learning experiments. A neural network model consisting of multiple modules is proposed, in which the interactive dynamics between the bottom-up perception and the top-down prediction are investigated. Our experiments with a real mobile robot showed that the incremental learning of the robot switches spontaneously between steady and unsteady phases. In the steady phase, the top-down prediction for the bottom-up perception works well when coherence is achieved between the internal and the environmental dynamics. In the unsteady phase, conflicts arise between the bottom-up perception and the top-down prediction; the coherence is lost, and a chaotic attractor is observed in the internal neural dynamics. By investigating possible analogies between this result and the phenomenological literature on the "self", we draw the conclusions that (1) the structure of the "self" corresponds to the "open dynamic structure" which is characterized by co-existence of stability in terms of goal-directedness and instability caused by embodiment; (2) the open dynamic structure causes the system's spontaneous transition to the unsteady phase where the "self" becomes aware.

1 Introduction

One of the crucial problems in consciousness studies is that terms such as "consciousness" and "self", which ought to describe subjective human experiences, are rarely defined in an objective and scientific manner. Dennett (Dennett, 1991) made this point when he wrote that "every author who has written about consciousness has made what we might call the first-person-plural presumption." In considering this problem, this paper explores the possibility

*I would like to thank Joseph Goguen for discussion and encouragement with this study from its inception. I also would like to thank the journal's referees who suggest various ideas to improve the contents.

that the constructivist approach using the power of system analysis reinforces some theories about the "self" obtained from phenomenological observations.

By the constuctivist approach, what we mean here is an approach to understanding various cognitive and behavioral functions of humans and animals by modelling them with artificial systems –i.e. building computer simulators, robot systems etc. Although the constructivist approach in the field of artificial intelligence, artificial neural networks and machine learning has made a contribution towards the understanding of general cognitive mechanisms such as recognition, learning or planning, it has not made a great contribution towards understanding the problems of consciousness. Although some might be able to build their own models of consciousness by embodying them with some "conscious robots", such attempts would just end up with Dennett's (Dennett, 1991) problem of "the first-person- plural presumption" again. Once Thomas Metzinger (Metzinger, 1998) argued that the problems of consciousness cannot be solved if people focus on only those problems, but they can be naturally resolved when mutual relations among other elements of human cognition are well understood. This argument can be extended further to say that consciousness or self might be seen in *a hermeneutic process* which appears in the iterative interaction among elemental cognitive processes, as Tsuda discussed (Tsuda, 1984.) This interpretation is supported also by Varela (Varela et al., 1991) who claims that consciousness is not self-existential but appears as a co-dependent structure among others. In this sense, the constructivists may not be able to design or model functions of consciousness or self directly but some analogies for them may appear in terms of structures and regularities in the interactions among cognitive and behavioural processes that are properly configured.

Varela argued in his book "the embodied mind" (Varela et al., 1991) that the essence of the embodied mind resides in the structural coupling which is established in the interaction between the bottom-up process originated from the objective world and the top-down process originated from the subjective mind. In our study, we attempt to reconstruct this sort of interactive process between the subjective mind and the objective world in a robot platform by which we examine what sort of structures can emerge from such interactions. (Here the word "the subjective mind" is used only metaphorically in our robot studies by meaning a process that attempts to interpret the "objective world" by means of its accumulated sensory-motor experience.) If certain structures are observed in our experimental results, we investigate whether the structure corresponds with any phenomenological observations that have been obtained in other disciplines. Although it might be true that the descriptions of the "self" or "consciousness" from the constructivist side can never be more than metaphorical, we expect that such comparison with the phenomenological observations can serve to strengthen the metaphors for connecting aspects of robot and human behaviours more closely. The following subsections will discuss our framework in details.

1.1 Puzzles involving "bottom-up" and "top-down" in robotics

For more than four decades, robots have been good experimental platforms for constructivist research investigating human mechanisms of intelligence and cognition. The first biologically inspired robot was developed by Grey Walter as shown in his book "the living brain" (Walter, 1952) in the 1950's. Walter describes robots built on cybernetic control principles which demonstrated goal-seeking behaviour, learning capability, and Ashby's (Ashby, 1952) idea

of homeostasis. His "turtle" robot, which was equipped with a simple cybernetic controller using optical sensors, exhibited quite complex navigation trajectories with avoiding obstacles and approaching light sources in the environment. This illustrated well that even simple control functions can generate quite complex behaviours when interacting with the environment.

From the 1970's to the 1980s, robots were often used as platforms for Artificial Intelligence (AI) research. Typical research objectives in those days were to study how to represent abstract models of the world and how to build action plans efficiently using the acquired representation of the world. In these studies, researchers believed that the most essential feature of intelligence was the logical manipulation of symbols. The studies of Shakey (Nilsson, 1984) at SRI and CART (Moravec, 1982) in Stanford represent well such research attitudes. Although this type of research produced rich results for the theories and methodologies of AI, the understanding obtained usually exhibited a poor performance when implemented in physical robots and tested in the real world. More specifically, the robot could not tolerate the discrepancies between the given model and its actual experiences in the real world; the robot could not plan its actions in real time.

So-called behaviour-based robotics were initiated by Rodney Brooks (Brooks, 1991) at the end of the 1980s in reaction to conventional AI research. The new research put less emphasis on the performance of higher order cognition such as abstraction, representation or planning. Instead, the focus was on the interactions between the robot and its environment at the sensory-motor level. A robot perceives the sensory input from the environment, and the motor actions are directly determined by means of reflex responses to the sensory input. As the result of the iterative sensory-motor interactions, certain structures and regularities emerge in the coupled dynamics between the robot and its environment (Beer, 1995). The goal of the adaptive behaviour is to adapt the internal controller such that the emergent structure and regularity sustain the robots inside their zones of viability (Meyer & Wilson, 1991). This claim is actually a re-introduction of the one by Ashby (Ashby, 1952), Walter (Walter, 1953), and Braitenberg (Braitenberg, 1984) in the cybernetic era. The main difference between the behaviour-based robotics in the 1980's and the turtle robot of Grey Walter in the 1950's is the usage of digital computers instead of vacuum tubes for the internal controllers. The microprocessor technology allowed researchers to build small self-contained robots, while implementing various sophisticated control and adaptation algorithms on them. Regardless of such sophistication, the old and new robotics share the same principle – the internal processes are naturally situated in their environment by means of their structural coupling via sensory-motor loops.

It can be said that robotics has fluctuated between the two extremes – putting strong emphasis either on the top-down process of the subjective mind or the bottom-up process from the objective world. The strong negation of "representation" in the behaviour-based robotics left no rooms for "subjective mind" by which the realities perceived from the "objective world" can be anticipated and interpreted in the top-down manner. On the other hand, it can be said that AI researchers have been underestimating the significance of realities which arise in a bottom-up fashion from the objective world. This underestimation of the difference between the abstract model in the subjective mind and the reality of the objective world causes the so-called symbol grounding problems which have been discussed by Harnad (Harnad, 1990). We consider that constructivists need to focus seriously on the

relation between the bottom-up process from the objective world and the top-down process from the subjective mind, since the problems of consciousness are very likely to be found in the middle of these two pathways (Varela et al., 1991).

1.2 Symbol grounding problems

A common approach for combining the bottom-up and top-down pathways is to use hybrid-type models which consist of a lower level processing the real world signals with analogue pattern-matching (often by using the connectionism technique) and a higher mental level which deals with symbolic manipulations of the abstract world models. These two levels are interfaced by a device, called a categorizer, by which patterns can be categorized into clusters represented by particular symbols. The patterns can then be projected back by indexing the symbols. Harnad considered that the categorizer device is a strong candidate for solving the symbol grounding problems (Harnad, 1990). He further argued that hybrid symbolic/analog models employing the categorizers can be scaled up to human capacity levels more readily than pure symbolic models or purely neurodynamics models (Harnad, 1993). This argument might be overly optimistic. In reality, such architectures could still suffer from symbol grounding problems. Let us describe such an example in autonomous robot research.

The example which we introduce is about the landmark-based navigation of mobile robots which has been studied by many robot researchers (Kuipers, 1987; Mataric, 1992). A typical mobile robot, which is equipped with simple range sensors, travels around a certain office environment while sensing the range reading of its surrounding environment. The continuous flow of the sensory image is categorized into one of several predefined landmark types such as a straight corridor, a corner, a branch or a room entrance. The upper level constructs a chain representation of landmark types by observing sequential outputs of the categorizer while the robot explores the environment. Such an internal map consists of nodes representing landmark types and arcs representing transitions between them. This representation takes exactly the same form as a symbolic representation known as finite state machine (FSM). (A FSM consists of a finite number of discrete states and their state transition rules.) Once the robot acquires the internal map of the environment, it becomes able to predict the next sensation of landmarks during its travel by looking at the next state transition in the FSM. When the actual perception of the landmark types matches the prediction, the robot proceeds to the prediction of the next landmark to be encountered. An illustrative description is shown in Figure 1 .

The problems take place when this matching process fails. The robot will become lost since the operation of the FSM halts upon receiving an illegal symbol/landmark type. This is the symbol grounding problem. Some may argue that the addition of certain exception handling procedures to the FSM operations will remedy the situation. For example, one could write a program such that when unexpected landmarks occur, the current state transition rules of the FSM could be modified suitably. However, we argue that such exception handling procedures could face another symbol grounding problem since there is no logical way of knowing whether unexpected landmarks appear by means of temporal noise or due to permanent changes in the environment.

When systems involve the bottom-up and the top-down pathways, the systems inevitably

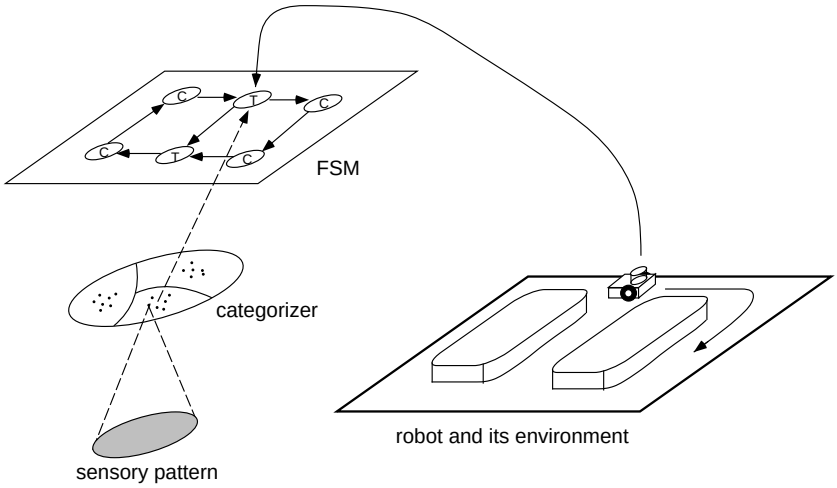

Figure 1: Landmark-based navigation of a robot using hybrid-type architecture consisting of FSM and categorizer.

encounter inconsistencies between the two pathways. The problem is how such inconsistencies can be treated internally without causing a fatal catastrophe in the system's operations. In hybrid systems, the higher symbolic levels, which are designed to resolve such inconsistencies, simply halt their operations upon facing a contradiction between the expectation of the FSM and the observation. We consider that both levels are dually responsible for any inconsistency and that the conflicts should be resolved through cooperative processes between the two levels. The cooperation entails iterative interactions between both sides through which optimal matching between the two sides is sought dynamically. The iterations in time are necessary not for the complete resolution of the conflict but for the postponement of the current inconsistency to a future time via which the systems can at least continue their consisting current operations. The major drawback of hybrid systems is that two pathways of the bottom-up analogue processes and the top-down symbolic processes cannot interact with each other intimately since the two pathways are defined in different metric spaces. In this paper, we propose an alternative – the dynamical systems approach (Kawato etal, 1990; Freeman, 1995; Smith & Thelen, 1994; Beer, 1995; van Gelder, 1998; Tani, 1996) – in which all systems including the mind, body and environment are defined as dynamical systems to enable their interaction to take place in the shared metric space.

1.3 Dynamical systems approach

The computational metaphors of dynamical systems have been discussed by many others (Crutchfield, 1989; Blum et al., 1989). It is, however, important to say that in dynamical systems we cannot recognize the separable entities "representation" and "manipulation" as one does in conventional computation. Ester Thelen and Linda Smith wrote in their recent book (Smith & Thelen, 1994) "Although behaviour appears rule-driven, there are no rules. There is a multiple, parallel and continuously dynamic interplay of perception and action, and a system that, by its thermodynamic nature, seeks a certain stable nature." In the

dynamical systems view, all that exists is the dynamical structures of the system and the resulting system's behaviour. The dynamical structure corresponds to the configuration of the vector flows in the phase space of the system. For example, if all vectors in the phase space of a system converge onto a point, the system behaves as characterized by equilibrium fixed point dynamics. If all the vectors converge onto a closed cycling orbit, the system behaviour is characterized by periodic oscillations known as "limit cycling dynamics". If the configuration of the vector flows becomes more complex and satisfies certain conditions, the system's behaviour is characterized by randomness and nondeterminism which is known as "chaotic dynamics". The new question is from where all the dynamical structures originate. We consider that the dynamical structures of the system are not something to be given or to be designed, but they emerge as self-organized through the iterations of the system itself. The dynamical structures generate the iterations of the system which, in turn, modify the original dynamical structure. The dynamical structure for cognition is established through such self- referential processes.

A difficulty exists in the embodiment of the higher cognitive levels by means of the dynamical systems approach. The question is how the dynamical systems approach can embody the complex subjective processes without employing the AI scheme of symbolic representation and manipulation. The key to solve the problem can be found in a recent scheme developed in the field of artificial neural networks, called recurrent neural network (RNN) learning (Elman, 1990; Pollack, 1991). The RNNs are considered to be adaptive dynamical systems from which dynamical structures can be tuned by means of neural connective weight modification using certain self-learning schemes. Elman (Elman, 1990) and Pollack (Pollack, 1991) showed that the RNNs can learn certain language syntactic structures from example sentences. The particular finding in their research is that grammatical rules cannot be seen explicitly in the neural internal representation, but the rules are actually embedded in attractor dynamics of the RNNs. Using the RNNs is suitable for our objective since we can exclude the "homunculus" from the systems which attempts to look down at the representation from the top and to manipulate the elements of the representation. The symbol grounding problem may not exist for RNNs since there exist no explicit forms for the symbols in the RNNs which need to be grounded.

Our cognitive architecture, which will be described in this paper, employs the RNN scheme for embodying the top-down subjective processes of the robot. We expect that intimate interactions will take place between the top-down processes of the RNN and the bottom-up pattern matching processes of the other conventional neural nets, since the processes share the same metric space comprising the real number system. In the experiments, we investigate how the total dynamical system evolves through the iterative interactions between the internal neural systems and the environment, focusing especially on the bottom-up and the top-down issues.

We will show our finding of specific dynamical phenomena from the results of iterated experiments and discuss its analogy to the notion of "self" found in the literature of phenomenology. Finally in this paper, we will describe our model of the "structure of the self" which is obtained by extending Strawson (Strawson, 1997) and Hayward (Hayward, 1998) studies about discontinuous occurrences of the "selves".

2 The Robot and Its Cognitive Tasks

Our experimental investigations concern the navigation learning of a mobile robot. For the experiment, we built a small mobile robot which is equipped with vision and tactile sensors as shown in Fig 2. The robot controls camera targeting, both horizontally and vertically, in addition to the manoeuvring of its wheels. We will now describe the task setting more specifically.

2.1 The task

The robot travels clockwise around a closed workspace following its outside wall as shown in Fig 3. The robot has to switch its attention between two visual tasks: wall edge following (for detecting the configuration of the wall on the left-hand side) and object recognition (for coloured objects appearing on the right-hand side of the robot). These two tasks are alternated between while the robot travels smoothly along the detected wall at a constant speed. The robot encounters two types of landmarks in the workspace: objects with coloured patterns on their surface, and corners in the wall. As the robot sequentially encounters these landmarks while travelling around the same workspace, the landmark sequences are learned via consolidation into a long term episodic memory. The robot then becomes able to anticipate which landmark will be encountered and when it will do so from the episodic memory. The recognition of a landmark proceeds as a cooperative process between the anticipation in the higher level and the perception in the lower level. The anticipation also activates the robot's attention towards a visual target. The visual attention is switched to searching for coloured objects on the right hand side of the robot when a coloured object is expected to be encountered in the near future. Otherwise, the vision is engaged in the wall edge following task. It is considered that the robot becomes situated to the environment when the anticipation and the perception match well for the event sequences. This navigation learning experiment was conducted in both a constant and an inconstant environment.

2.2 Cognitive problems in the task

The tasks described here are not trivial since the robot must perform the tasks utilizing only the finite cognitive resources available. What neural networks can do are always constrained in time and space. For example, the recognition of a visual object requires a certain convergence time for the network since a neuron's activation evolves according to its inherent time constant. The learning capacity is limited since the number of synaptic connections is finite. Moreover, the physical embodiment constrains the robot's cognitive processes significantly. The robot's vision can only see within a finite range in the workspace at one moment. The visual attention switch takes a nonzero time in order to rotate the motors of the camera head. The difficulty is that the robot has to interact with the world in a real time manner under various cognitive and physical constraints. This difficulty is explained more specifically in the following.

- The time delays in the visual recognition and the attention switch make the visual processes of the robot difficult. Problems take place when the robot travels in the workspace while alternating its visual attention between wall-following and object

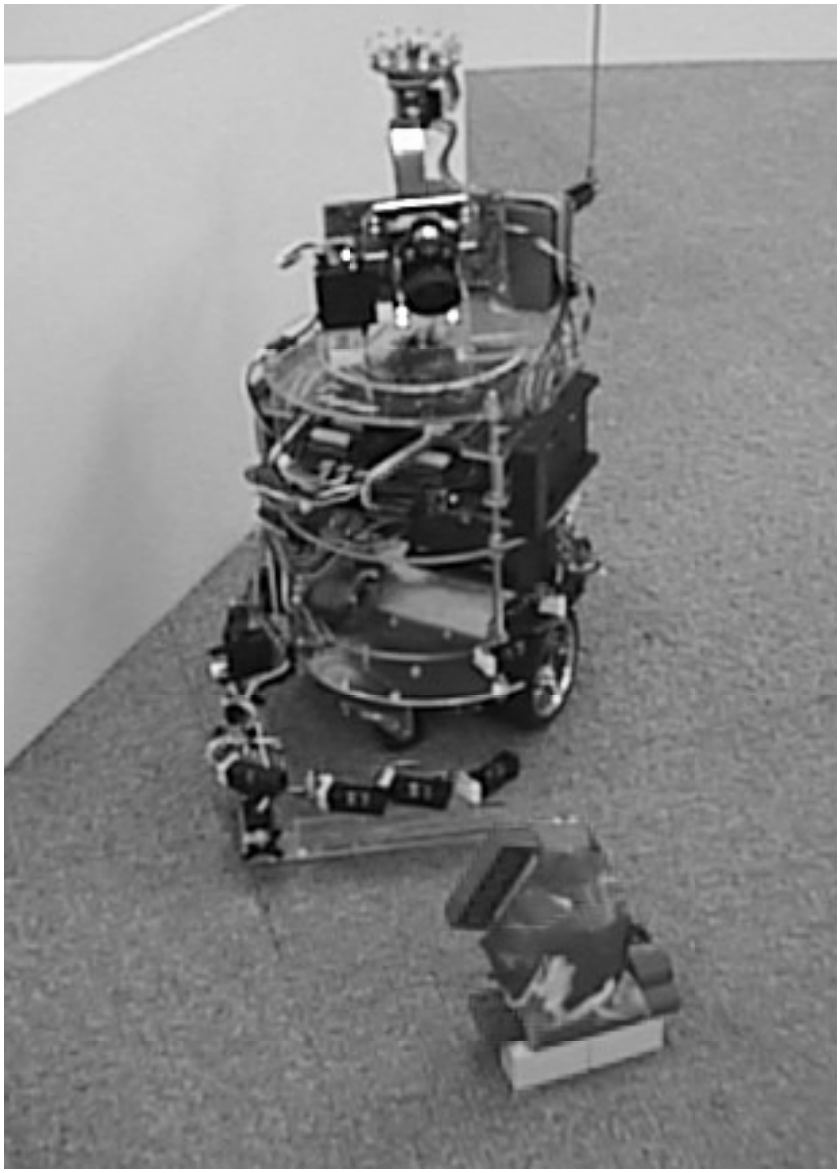

Figure 2: The vision-based mobile robot used in the experiments.

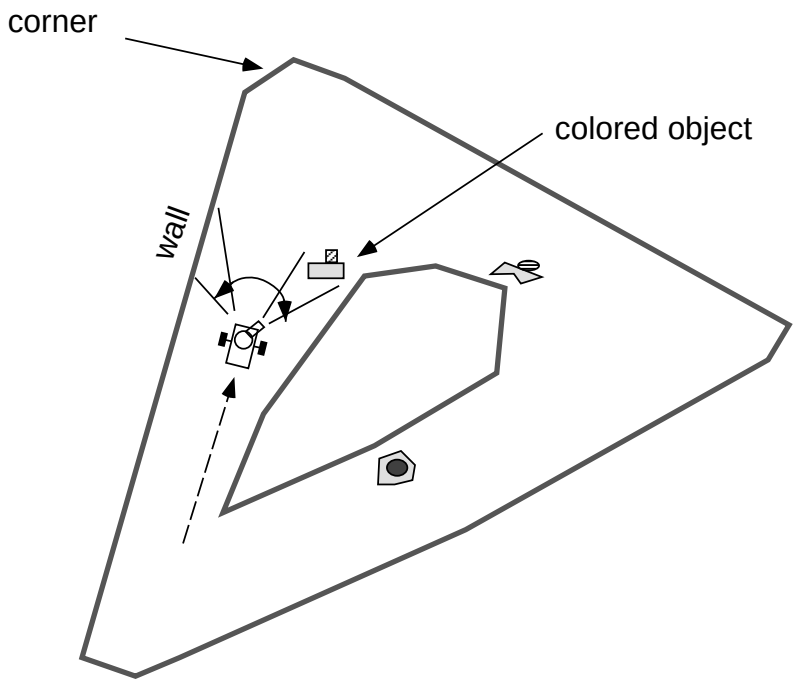

Figure 3: Robot navigation with landmark objects and corners.

recognition. If attention is engaged in the recognition of a colored object for too longer, the robot might overlook the wall collide with it. On the other hand, if object recognition is engaged for only a shorter period, the recognition process could result in mis-identification of the object since the process might be terminated before complete convergence.

- The cooperation scheme between the top-down anticipation and the bottom-up perception is a non-trivial problem. The correct top-down anticipation could stabilize the lower perception processes. However, if the learning of the event sequences is insufficient, the resulting incorrect anticipation may interfere badly with the perception of the next event in the lower level. Strong top-down anticipation could override the bottom-up perception, which might cause "illusions". The problem is that it is hard to determine in what ratio the top-down and the bottom-up processes should contribute to the recognition processes.

- The incremental learning scheme in the inconstant environment faces a dilemma between stability and plasticity since learning new experiences often interferes with what was learned in the past. When the new experience and the former memory conflict with each other, the conflict cannot be resolved easily since the robot itself cannot judge on this occasion whether it should acquire the new experiences as being the correct one and forget the former memory as being obsolete, or to ignore the new experience as being a misleading result and to preserve the former memory.

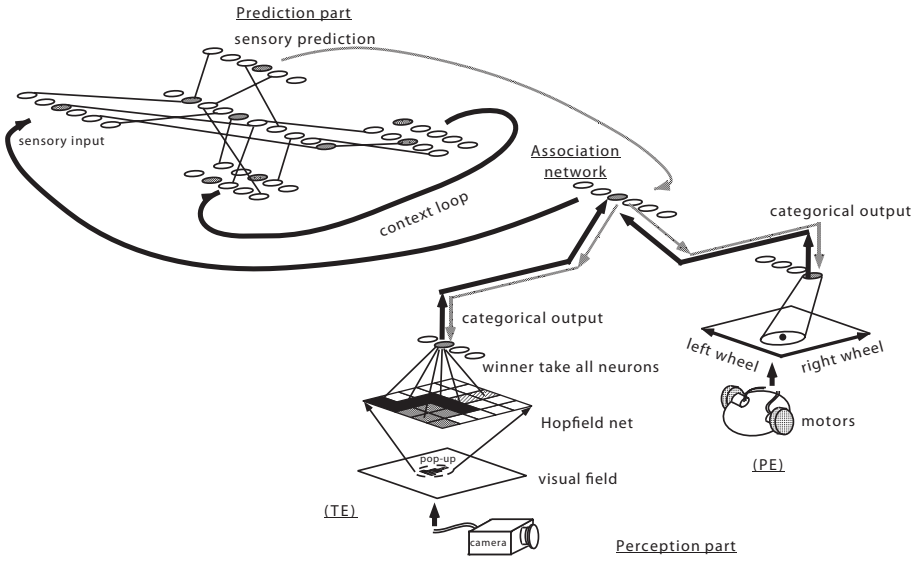

Figure 4: Proposed architecture consisting of multiple neural networks.

3 Models

This section introduces the cognitive architecture employed in our robot. The architecture consists of several neural network modules in which some aspects of the modelling are based on physiological findings shown in the literature, but only in quite an abstract manner. The architecture is, rather, built to reflect our interpretation of how the bottom-up and the top-down interactions might take place in the cognitive processes of animals or humans. In the following, we first describe our modelling procedure, including the neural architecture, the learning scheme and the arbitration mechanism between the top-down and the bottom-up pathways, in a very abstract manner in order that non-technical readers can understand the content. The later subsections will describe the details which non-technical readers may like to skip.

Figure 4 shows a schematic diagram of our proposed neural net (NN) architecture consisting of multiple modules. The architecture consists of three parts, namely the perception part, the association part and the prediction part. The perception part consists of the PE module and the TE module which recognize the currently object of visual focus (a landmark) in terms of a "where and what" manner (Ungerleider & Mishkin, 1982). (PE and TE are actually the names of specific regions in the parietral lobe and the inferior tempral lobe in the human neocortex.) Two streams of neural activities of "where and what" are sent to the association part where they are represented as integrated neural activities.

The associated neural activities corresponding to the current visual object are also sent to the top-down prediction part where a prediction of the next landmark to be encountered is generated. This prediction in terms of "where and what" is sent back to the perception part through the association part so that recognition can be carried out by taking account of the top-down prediction. The neuro-dynamics with two input forces from the top-down prediction of "what" and the bottom-up perception determine the actual recognition outcomes. The prediction in terms of "where" is used to direct the visual attention of the robot

to the direction expected for the next landmark encountering.

The learning in both the perception part and the association part is conducted each time the stimulus comes. The learning in the prediction part is conducted incrementally, which models the consolidation processes of memory transfer from short-term to long-term that Squire (Squire et al., 1984) observed in some animals and humans.

The architecture employs another adaptation mechanism that arbitrates the interaction between the top-down prediction and the bottom-up perception. In this mechanism, the top-down prediction is more biased when the predictability becomes better. On the other hand, the bottom-up perception is more biased with allocating longer time for the visual perception when the predictability gets worse.

3.1 Neural Net Architecture

In this subsection each module in the proposed neural architecture is described in detail.

(a) The perception part: In the perception part, the TE module receives the visual image from a video camera and the PE module receives the dead-reckoning vector from the rotation encoder of the wheels.

When a coloured object is detected in the visual field during the object search, the object is foveated by targeting the camera head towards it. The region of colour "pops-up" and pixels in the region are sent to an associative memory implemented using a Hopfield network (Hopfield & Tank, 1985) in the TE module. The Hopfield network consists of 10x10x3 neurons corresponding to the 10x10 pixel size with three-colour information for each pixel. The Hopfield network can memorize the pixel patterns by associative learning using the pixels through iterative exposures to the patterns. It is well known that each memory pattern is stored in a corresponding fixed point in the hypothetical potential field defined in the network (Hopfield, 1986). The Hopfield network is bi-directionally connected to an array of winner-takes-all neurons (Amari & Arbib, 1977; Waugh & Westervelt, 1993). Each neuron in the winner-take-all array is inhibitively connected. Each neuron also receives the top-down prediction inputs from the integration part. In the recognition process, the neural activation of the Hopfield net and the winner-takes-all neurons are dynamically computed simultaneously upon receiving the continuous video image from the bottom and the landmark prediction from the top. The neural state of the Hopfield net converges onto a fixed point attractor in the potential field and at the same time each neuron in the winner-take-all neuron array competes to be a winner. The winning neuron represents the categorical output of "what" in this module. Convergence takes up to several seconds in the real implementation using the robot. After convergence, associative learning is conducted for the internal connective weights in the Hopfield network as well as for the cross-connective weights between the Hopfield network and the winner-takes-all network so that all the connections among the activated neurons can be reinforced.

In the PE module, the winner-takes-all neurons dynamically compute the categorical output representing "where" by receiving the dead-reckoning data from the wheels and the top-down prediction from the integral part. The dead-reckoning data represents the rotation sum of the left and right wheels starting from the previous encounter with a landmark to the current encounter. Therefore, the categorical outputs of "where" represent the movement of the robot relative to the position of the previous landmark encountered.

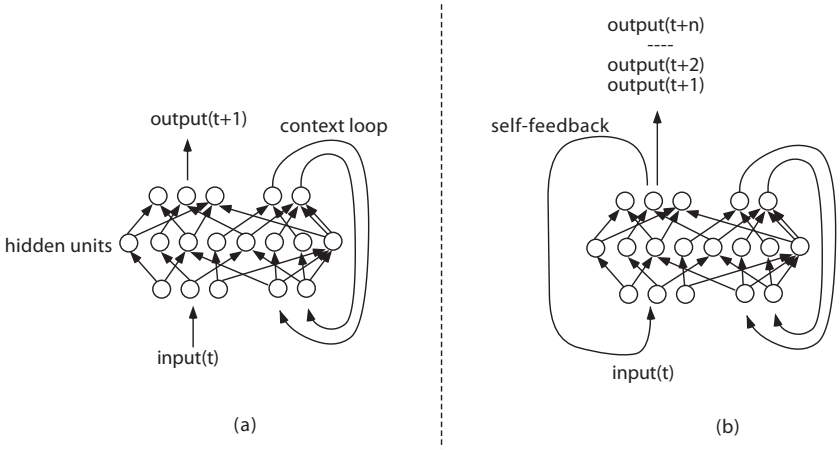

Figure 5: The RNNs operated in (a) the open loop for the one-step prediction and in (b) the closed loop for the lookahead prediction.

(b) The association part: The association part receives categorical outputs from both the TE and PE modules. When the association network receives a new combination of categorical outputs from the TE and PE modules, a neuron is allocated and positive connections are made to this neuron from each winning neuron in the TE and PE modules. In this manner, a single neuron activated in the association network represents the association of two categorical outputs of "what" and "where". The activity of top-down prediction from the higher level propagates through these established connections to both the TE and PE modules.

(c) The prediction part: The prediction part in our architecture may correspond to the prefrontal cortex in a primate. Higher order cognitive functions such as working memory, prediction and planning have been widely studied in primates (Fuster, 1989). Some researchers consider that such episodic memory of event sequences are stored in the so-called place-cells (O'Keefe & Nadel, 1978) in the hippocampus rather than the prefrontal cortex, as a large body of lesion studies shows evidence of hippocampal involvement in spatial cognition and episodic memory in rodents (O'Keefe & Nadel, 1978; Wilson, 1994.) It is, however, still unclear whether this hippocampus theory can be applied to monkey or human cases and if the theory accounts not only for the short-term memory but also for the long-term memory. Therefore, what we assume in the current model is that the hippocampus stores the episodic memory of event sequences only for temporary memory, and these sequences are incrementally transferred to the prefrontal cortex in order to form the long-term memory through the so-called consolidation process (Squire et al., 1984). Then we assume that the actual prediction of the next coming event sequence is conducted by using this long-term memory which is embodied by a RNN in the current model. A later section will describe more details about the implemented consolidation mechanism.

In the proposed architecture, a conventional RNN, receiving its current activation state, (Elman, 1990; Jordan & Rumelhart, 1992; Pollack, 1991) learns to predict the activation in the association part due to the next landmark encounter. Figure 5 shows details of the RNNs operating in open-loop mode and in closed-loop mode. Upon receiving the inputs, the

activation of the output units is calculated by means of nonlinear mapping using the synaptic weights shown in Fig. 5(a), where the outputs represent the prediction. This operation mode of the RNN is called the open-loop mode. We employ Jordan's idea of context re-entry which enables the network to represent the internal memory (Jordan & Rumelhart, 1992). The current context input is a copy of the previous context output: by this means the context units remember the previous internal state. The navigation problem is an example of a so-called "hidden state problem": a given sensory input does not always represent a unique situation/position of the robot. (For example, quite similar shapes of corners or objects can be observed from different locations.) Therefore the current situation/position is uniquely identifiable not by the current sensory input, but by the memory of the sensory-motor sequence stored during travel. Such memory structure using the context re-entry is self-organized through the learning process.

The RNN can function as an autonomous dynamical system by installing a closed loop from the outputs to the inputs as shown in Fig. 5(b). The prediction for the next step obtained from the current output units is re-entered into the input units for the next step of computation. By iterating the dynamical map with the closed loop, the RNN can conduct lookahead prediction for input sequences of arbitrary length, which resembles the performance of memory rehearsal. This iteration of the RNN with the closed loop can exhibit various dynamical structures depending on the connective weights developed though learning (Tani & Fukumura, 1995). This is similar to the logistic map (Devaney, 1989) which can exhibit bifurcations from a fixed point, limit cycling, or chaotic dynamics depending on its initial parameters. In the later sections, it will become apparent that the understanding of such dynamical structures within the RNN is essential to the current study.

The RNN used in our experiment has 9 input units, 9 output units, 25 context units and 25 hidden units. The input units adopt a discrete local representation (only one unit is set to 1.0; the others are set to 0.0 which corresponds to the winner in the association part), while the context units and the hidden units are activated to take a grey distribution of values. The output units are also activated to take on grey values, where a high activation state of an output unit denotes a high confidence that the corresponding neuron in the integration part is to be activated in the next event step. The grey valued top-down prediction by the output units is sent to both of the TE and the PE module through the association part. The expectation of landmarks in the form of "what" and "where" is utilized for the visual attention switch.

3.2 Visual targeting

The two visual tasks of wall following and object recognition are alternated between according to the top-down prediction obtained in the neural net architecture described in the previous section. In the following, details of the visual targeting mechanism for these two task modes are described.

The basic mechanism for camera targeting is realized by a hand-coded program using the assumption that the walls and the coloured objects appear on the left-hand side and on the right-hand side of the image space respectively. In the wall following task, the camera head turns maximally to the left and focuses on the edge between the wall and the floor. The camera head then turns gradually to the forward direction, following the perceived edge line

as foveated in the centre of the visual field. The measured trajectories of the head's rotation in the horizontal and vertical directions represent the shape of the wall edge. This single movement of the camera head from the extreme left to the forward direction takes about 2 seconds. The current motor commands for the wheels are determined by a hand-coded function to enable the robot to travel smoothly, thus avoiding collisions with the wall.

In the object search and recognition, the camera head sweeps from the frontal side to the right-hand side in searching for a coloured object. Once a coloured object is targeted, the vision recognition process is initiated. One important note is that the robot cannot manoeuvre its wheels while it engages in the object search and recognition task. The robot risks of collisions with the walls during this engagement. The next section will describe how attention is switched between these two visual tasks using the top-down prediction.

3.3 Arbitration between the top-down and the bottom-up

As we have described in the previous section, the visual recognition processes face two essential problems concerning (1) the conflicts between the top-down and the bottom-up processes and (2) the time delay. While the robot is new to the workspace with a small amount of learning experience, the predictions of the RNN are vague and erroneous. Therefore, an erroneous prediction can lead to fluctuations in the timing of the attention switch; erroneous prediction might also interfere with the bottom-up perception in the Hopfield net. Moreover, under these circumstances convergence of the visual recognition takes a longer time, since the pressure from the top-down prediction is weak and the attractors organized in the Hopfield network are shallow. On the other hand, when the robot becomes familiar with the workspace, more accurate prediction enables correct timing of the attention switch for the upcoming landmarks; accurate prediction also enables the bottom-up perception processes in the Hopfield net to be stabilized more rapidly.

In this situation, what is lacking in this cognitive architecture is a type of "awareness" property, which one means that the robot can be aware of its own state of familiarity or "situatedness" in the environment in which it is currently engaged. Such a self-measured state can be utilized internally to arbitrate between the top-down and the bottom-up processes. For this purpose, the predictability measure in the network is introduced, which represents the degree of agreement between the top-down expectation and the bottom-up perception. (The measure is the sum of the dot products between the two vectors corresponding to the prediction and the outcome for each TE and PE module.) The following arbitration scheme is considered using this predictability measure.

1. When the predictability measure is more than a certain threshold value, the visual attention is switched to the right-hand side in preparation for object recognition by following the top-down prediction. On the other hand, when the predictability measure is less than a certain threshold value, attention is switched alternatively between the wall following on the left-hand side and searching for the objects on the right-hand side, ignoring the top-down prediction.

2. The strength of the top-down prediction with respect to the winner-takes-all neurons in the TE module and in the PE module is modulated in proportional to the predictability

measure. Therefore, the actual top-down input given to the winner-takes-all neurons is obtained from the prediction value from the RNN multiplied by this strength.

3. The maximum iteration time allocated for the convergence dynamics in the Hopfield net is modulated in inverse proportion to the predictability measure.

These schemes mean that the robot tends to rely on the bottom-up process, ignoring the top-down process if the robot becomes suspicious about its own prediction reliability; otherwise the robot tends to be relatively unreliant on the bottom-up process, being dependent primarily on the top-down process.

3.4 Incremental learning and consolidation process

It is difficult for RNNs to learn incrementally the information received. It is generally observed that the contents of the current memory are severely damaged if the RNN attempts to learn a new teaching sequence. One way to avoid this problem is to save all the past teaching data in a database. When new data is received, it is added to the former data in the database, and all the data is then used to re-train the network. Although this procedure may work well, it is not biologically plausible.

Observations in biology show that some animals and humans may use the hippocampus for temporary storage of episodic memories (Squire et al., 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. Recent experiments (Wilson, 1994) on the hippocampal place-cells of rats show evidence that these cells reinstate the information acquired during daytime active behavior. McClelland (McClelland et al, 1994) further assumes 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 apply these hypotheses to our model of RNN learning. In our system, the sequence of events experienced, which may correspond to a temporary episodic memory, is stored in the "hippocampal" database. In the consolidation process, the RNN which corresponds to the prefrontal cortex rehearses the stored memory patterns. This rehearsal can be performed in the closed-loop mode, as described in the previous section. The sequential patterns generated by the rehearsal are sent to the hippocampal database. The RNN can be trained using both the rehearsed sequential patterns which correspond to the former memory and the current sequence of new experience, by using the back-propagation through time learning algorithm (Rumelhart, 1986). The re-training of the RNN is conducted by updating the connective weights obtained in the previous training.

4 Experiment

The learning experiments were conducted in the "original workspace" and in the "modified workspace" shown in Figure 6 (a) and (b), respectively. Five landmarks consisting of two visual objects and three wall corners exist in the original workspace. The robot navigates until it encounters 15 landmark steps during each travel. After each travel, the robot "sleeps"

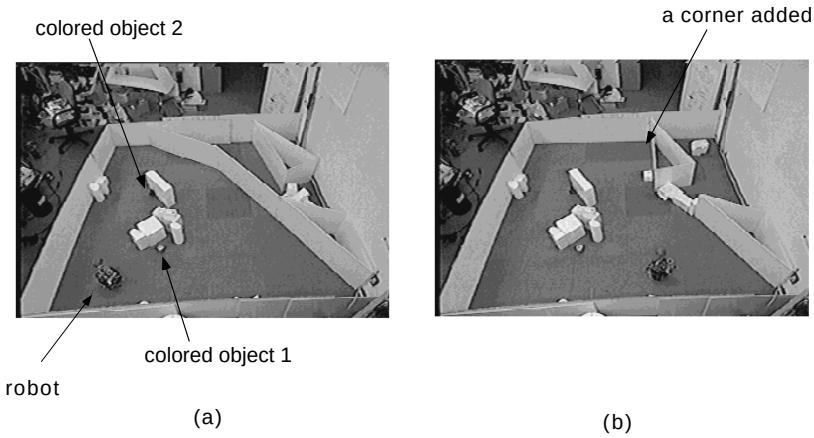

Figure 6: (a) the original workspace and (b) the modified workspace used for the learning experiment.

to produce the memory consolidation process as has been described in the previous section, in which the RNN is trained for 6000 learning steps. After sleeping, the robot starts to travel again; the training process is terminated. This cycle involving travel and learning is repeated.

In this experiment we focus on how the internal dynamical structure of the system evolves in the course of incremental learning and also on how the interactions evolve between the internal and the environmental dynamics. For this purpose we observed the evolutionary paths of the RNN dynamics and of the prediction error. The evolutionary path of the RNN dynamics is observed by means of plotting the bifurcation diagram of the RNN dynamics during the learning process as well as by making phase plots of the RNN dynamics at particular instants in time. The prediction error is given by as 1 subtracted by the normalized predictability as defined in the previous section. In the following subsections, we describe two experiments: the adaptation to the original workspace and the re-adaptation to the modified workspace. We devoted the last sub-section for the brief summary of the experimental results. Non-technical readers may skip the following two sub-sections and may read only this summary in order to understand the basic results.

4.1 Adaptation to the original environment

Three independent trials of adaptation to the original environment were conducted. In each trial, the cycle involving travel and learning was repeated 7 times. Figure 7 to 9 show the prediction error at each landmark encounter, the bifurcation diagram of the RNN dynamics and the phase plot at a particular time for each trial.

The bifurcation diagram of the RNN shows the evolutionary path of the closed loop RNN dynamics for each learning period during “sleep”. For example, the plotted values from the 0th learning period to the 1st learning period in the diagram show how the activation state of the RNN evolves due to the learning of the landmark sequence experienced during the first 15 steps. We plotted the average of the context neuron’s activation state generated by each iteration of the closed-loop dynamics of the RNN. We observed how this activation

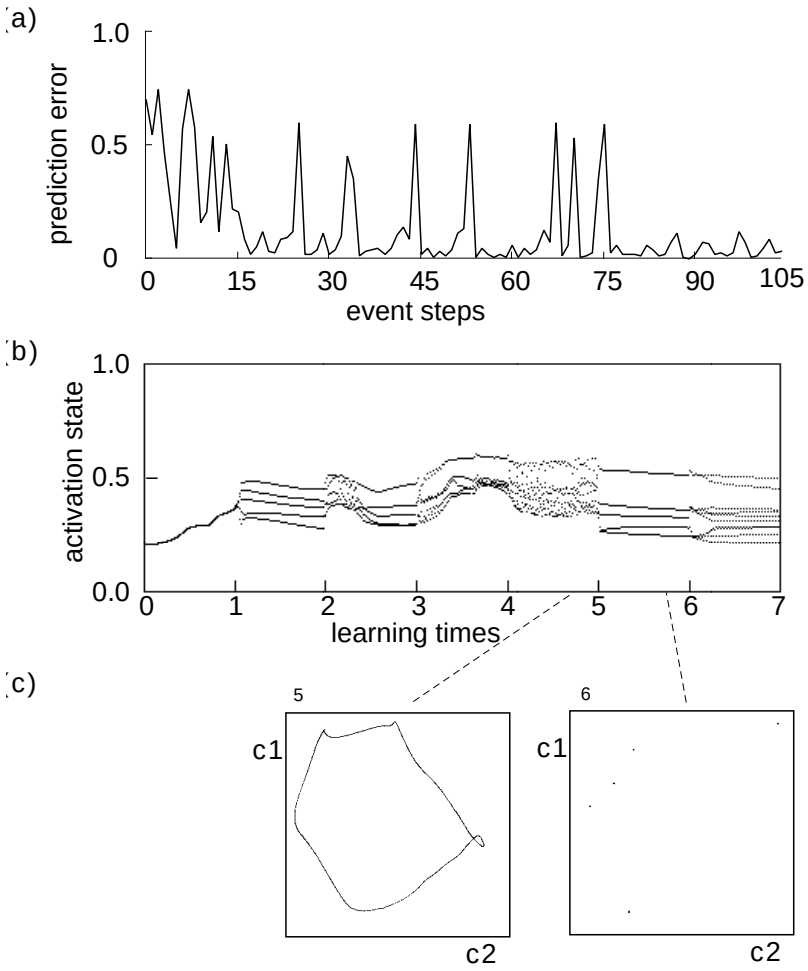

Figure 7: Trial-1: (a) the prediction error, (b) the bifurcation diagram of the RNN dynamics and (c) the phase plots at particular times. The times are indicated by the dashed lines.

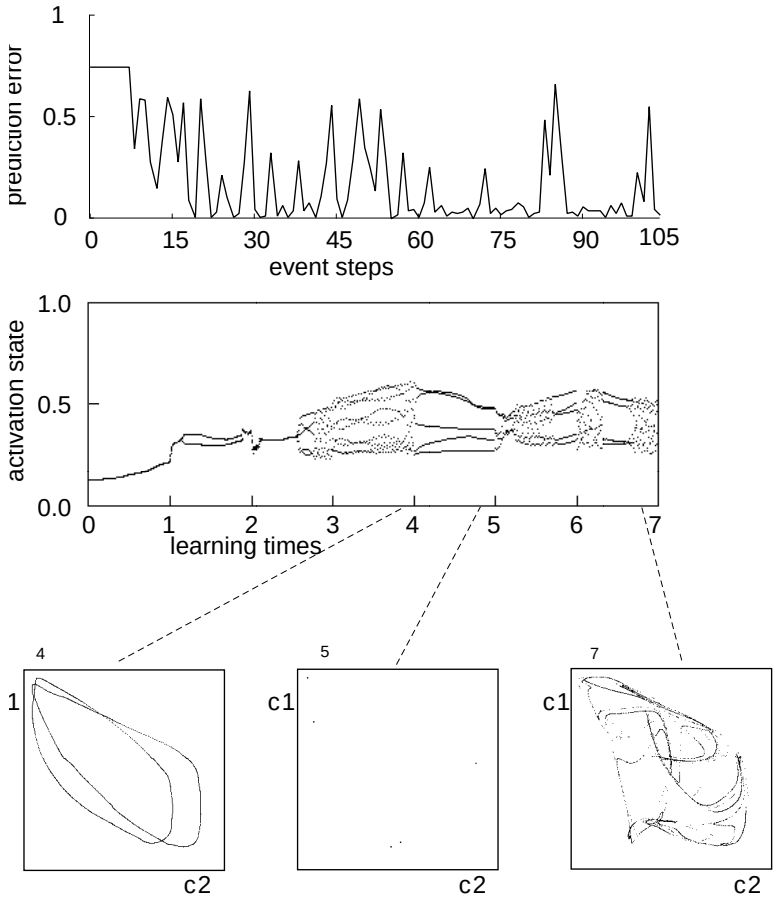

Figure 8: Trial-2: (a) the prediction error, (b) the bifurcation diagram of the RNN dynamics and (c) the phase plots at particular times. The times are indicated by the dashed lines.

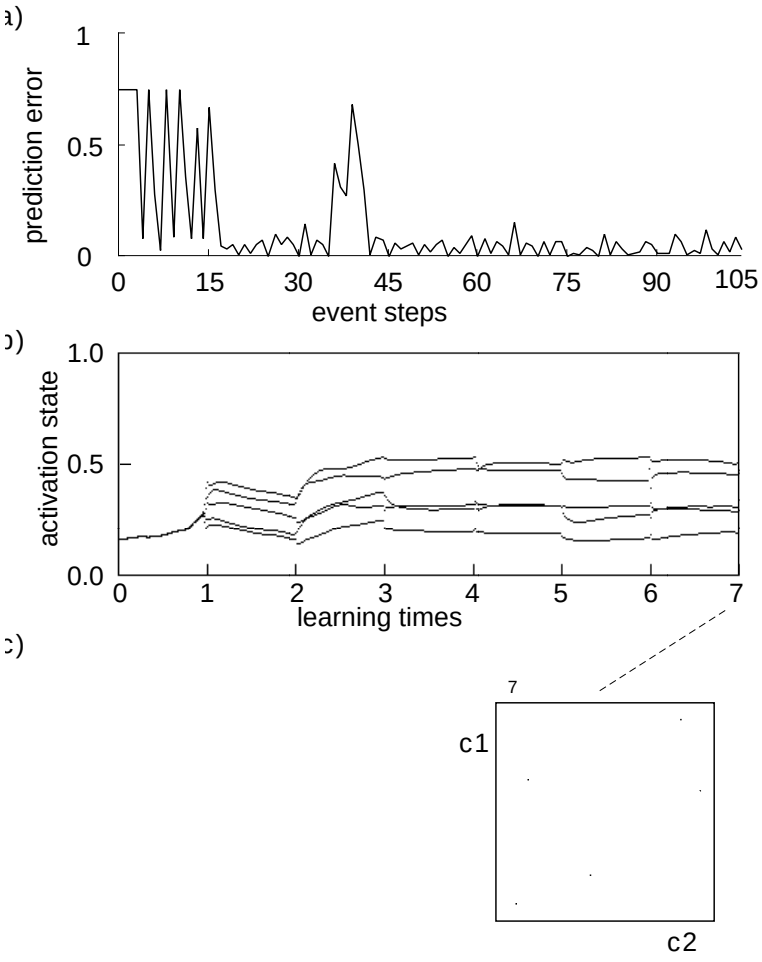

Figure 9: Trial-3: (a) the prediction error, (b) the bifurcation diagram of the RNN dynamics and (c) the phase plots at particular times. The times are indicated by the dashed lines.

state changed while the connective weights gradually changed due to the learning process [1]. The phase plot was drawn using the connective weights sampled in the particular learning period in order to show snap-shots of the evolving attractor. We plotted the c_1 and c_2 values which are average activation over one half of the context neurons and that over the other half of the context neurons, respectively. The (c_1, c_2) points were plotted from the results of 6000 iterations of the closed loop RNN dynamics using the sampled connective weights.

In all three trials, the prediction error is quite high in the beginning because of the initially random connective weights. After the first learning period, the predictability is improved to a certain extent in all three trials but the errors are not minimized completely except in trial-3. Prediction failures take place intermittently during the course of the trials. From the bifurcation diagram, we can also see that the dynamical structure of the RNN varies from time to time in trial-1 and trial-2. In the first learning period, fixed point dynamics –i.e. converging onto one activation state – appear in all three trials. Thereafter, phase transitions take place during the second learning period for all trials; limit cycling dynamics with a periodicity of 5 appear in trial-1 and in trial-3, while dynamics with a periodicity of 2 appear in trial-2. We observe that limit cycling dynamics with a periodicity of 5 appear frequently in the course of the trials. The periodicity of 5 is significant since it corresponds to the 5 landmarks which the robot encounters in a circuit of the workspace. It should be noted that the observed limit cycling dynamics with a periodicity of 5 do not remain stationary. The periodicity of 5 disappears intermittently and other dynamical structures emerge. The phase plots show such emergent dynamical structures more clearly. We observe a non-periodic attractor in the 5th learning period in trial- 1 in the phase plot. The attractor was identified as being weakly chaotic since its Lyapunov exponent [2] (Devaney, 1989) was computed to be slightly positive at 0.0011. These weakly chaotic dynamics disappear soon after and the periodicity of 5, which corresponds to the 5 points in the phase plots, appears in the 6th learning period. In trial-2, we observed quasi-periodic dynamics, limit cycling dynamics with a periodicity of 5 and a strange chaotic attractor with a Lyapunov exponent of 0.14 in the 4th, 5th and 7th learning periods, respectively.

Based on the results, we see that there exist two distinct phases which are the steady phase represented by the limit cycling dynamics with a periodicity of 5 and the unsteady phase characterized by non-periodic dynamics. It is also seen that the shifts between two phases take place arbitrary in the course of the time development. In order to clarify the differences between these two phases, we selected two sequences, namely from the 60th step to the 74th step and from the 75th step to the 89th step, both in trial-1, which represent the unsteady phase and the steady phase, respectively. We compared the actual perception sequence and its prediction made by the RNN for the two sequences. Figure 10 shows such a comparison. In Fig. 10(a), we see that the prediction is inaccurate with the actual perception in some parts of the sequence while the prediction is highly accurate for the sequences in Fig. 10(b). A more important observation is that the actual perception sequence in these two sequences differs in some parts even though the robot travelled through the same workspace in both periods. We can see the periodicity of 5 in the actual perception sequence in Fig. 10(b)

[1] The bifurcation diagram is usually obtained by gradually changing a parameter in the target dynamical system. In the current case, the connective weights correspond to the system's parameters being changed.

[2] In dynamical systems theory, it is well-known that a positive, zero or negative Lyapunov exponent denotes chaos, quasi-periodic, and limit cycling and fixed point dynamics, respectively

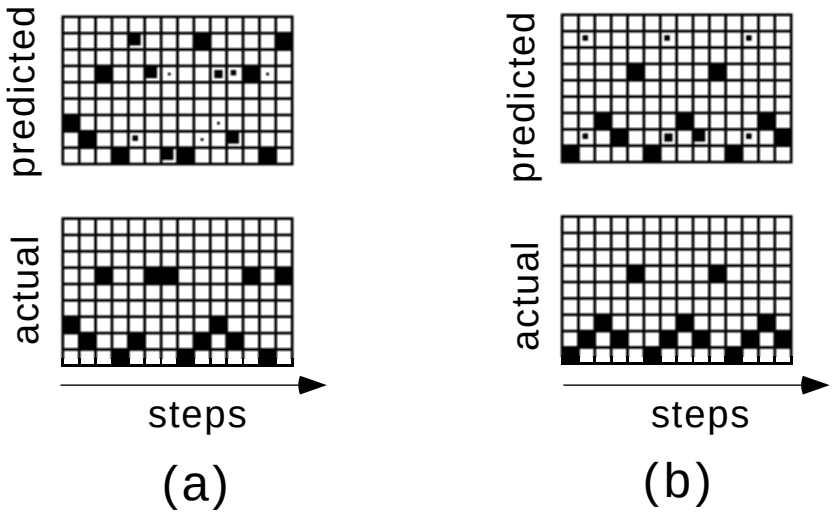

Figure 10: Comparison between the actual perception and its corresponding RNN prediction sequence (a) in the unsteady phase (60th step to 74th step) and (b) in the steady phase (75th step to 89th step), both in trial-1.

while such a periodicity cannot be seen in the sequence in Fig. 10(a). In order to elucidate this observation, we compare the actual robot trajectories observed in these two periods. Figure 11 shows the robot trajectories measured in these two periods by a camera mounted above the workspace. It is seen that the trajectory winds more in (a) than in (b) especially in the way objects or corners are approached. We infer that the maneuvering of the robot is more unstable in (a) because the robot spent a longer period on the visual recognition of objects due to the higher value of the prediction error. (Recall that the maximum iteration time allocated to visual recognition is inversely proportional to the predictability measure.) Therefore the robot took a higher risk of mis-detection of landmarks when its trajectory meandered during this period. In fact, the robot mis-detected corners and objects when its

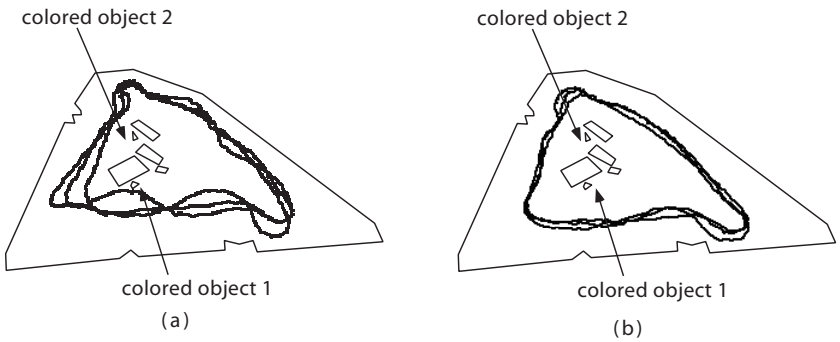

Figure 11: The robot trajectories measured (a) in the unsteady phase (60th step to 74th step) and (b) in the steady phase (75th step to 89th step).

trajectory meandered severely during this period. On the other hand, in the period from the 75th step to the 89th step, the detection sequence of landmarks became more deterministic and the robot travelled smoothly with greater prediction success.

An important comment on the above is that the observation of the steady and unsteady dynamics is attributed not only to the internal cognitive processes arising in the neural networks but also to the physical movements of the robot's body as it interacted with the external environment. It was observed that the change in the visual attention dynamics due to the change in predictability caused drifts in the robot's manoeuvring. These drifts resulted in mis-recognition of the upcoming landmarks, which led to the re-adaptation of the internal memory and a consequent change in the predictability. The dynamical interactions took place between all processes including attention, prediction, perception, learning and behaviour which led to a non-trivial time development of the total system.

4.2 Re-adaptation to the modified environment

When the the robot learning experiment in trial-3 reached to the 105th step, the original workspace was modified by adding one sharp corner as shown in Fig 6 (b). Thereafter, the re-adaptation experiment was conducted until reaching the 225th step was reached, which included 15 learning periods. Figure 12 shows the normalized prediction error at each landmark encounter, the bifurcation diagram of the RNN dynamics and the phase plot at particular times. In this figure, it can be seen that the prediction error becomes larger soon after the workspace is modified and this tendency continues until around the 165th step. The bifurcation diagram shows the appearance of non-periodic dynamics in the RNN during this period; the emergence of strange chaotic attractors is seen in the corresponding phase plots. We found that the strange attractors have positive Lyapunov exponents. The prediction error decreased after 11 learning periods and a periodicity of 6 was observed in the RNN closed-loop dynamics after 14 learning periods.

It was observed that the re-adaptation process is a non-trivial task for the robot, since the former memory based on limit cycling dynamics with a periodicity of 5 and the new experience with a periodicity of 6 conflict with each other in the learning processes. Immediately after the modification of the workspace, relatively strong top-down prediction based on the former memory tends to override the actual perception because the averaged predictability is still high. This destabilized the pre-learned associations between the actual sensory patterns received in the lower level modules and the corresponding structures in the RNN. In fact, the structures corresponding to the landmarks changed after the modification of the workspace – i.e., the RNN outputs corresponded differently to the images, in comparison with the period prior to the modification. Such reorganization of the association between the higher and the lower level representation increased the confusion of the system during the re-adaptation.

4.3 Summary of the experimental results

This subsection gives a summary of our experimental results. In the experiments with the original environment, the time-development of the system dynamics in the course of the incremental learning of the environment exhibits spontaneous switching between the steady

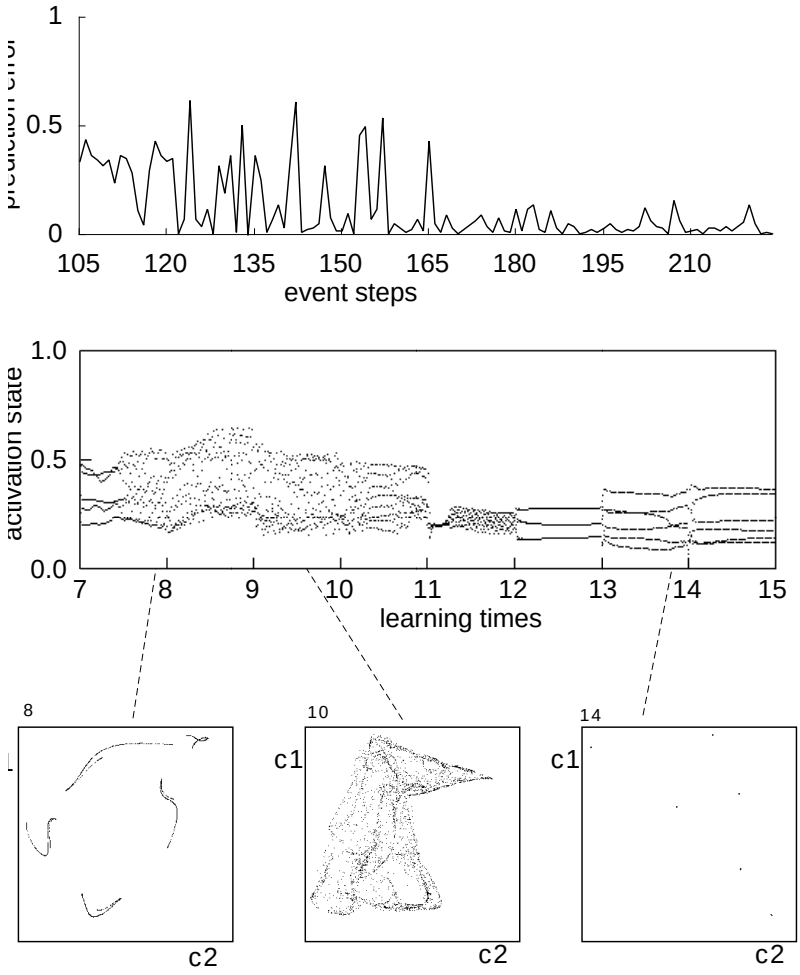

Figure 12: Trial-3 continued in the modified workspace: (a) the prediction error, (b) the bifurcation diagram of the RNN dynamics and (c) the phase plots at particular times. The times are indicated by the dashed lines.

phase and the unsteady phase. In the steady phase, the robot travels smoothly, showing good prediction of the coming landmark sequences where a simple periodicity is observed in the internal neural dynamics. This periodicity corresponds to the number of landmarks allocated in the workspace. On the other hand, the prediction goes wrong in the unsteady phase where non-periodic chaotic attractors are observed in the internal dynamics. The actual trajectory of the robot manoeuvring becomes much more unstable in the unsteady phase since much visual attention is directed to the landmark objects than to the walls to be followed. The experiment in the modified environment shows that the unsteady phase lasts for a while before the re-adaptation of the internal neural dynamics is achieved.

5 Analysis and Discussion

5.1 The open dynamic structure

Our experiment demonstrated that the internal dynamics of the RNN evolved by switching intermittently between limit cycle, quasi-periodic and chaotic dynamics. These fluctuations in the RNN learning are caused mainly because (a) the sequences to be learned contain a certain nondeterminism and (b) the learning process of the RNN behaves arbitrarily in determining its connective weights as has been shown by Ikegami and Taiji (Ikegami & Taiji, 1998). As described above, the RNN obtains nondeterministic sequences of landmark recognition events. This RNN learning involving finite sequences can result in various interpretations of the data obtained. Let us consider an example. Suppose that the RNN receives in one instance a sequence such as "a b c a d c a b c". The RNN learning process could interpret the rules underlying the sequence as being a simple periodic pattern of "a b c", ignoring the appearance of "d" in the second period. Other possibilities include a long periodic pattern of all the 9 letters in the sequence, or a period 3 sequence containing nondeterminism in that "a" can be followed by "b" or "d" in a nondeterministic fashion. The third case is especially interesting as it implies that the observed nondeterminism in the sequence is embedded in the deterministic chaos of the RNN dynamics. Mathematical and numerical analysis of this case is given in Ref. (Tani & Fukumura, 1995). An important remark is that the learning is somewhat arbitrary since all three of the above interpretations of the given sequence are possible. By this means, it can be said that our robot re-interprets the world as being deterministic or nondeterministic in an arbitrary manner after every learning process.

The dynamical structure observed in the system can be characterized as shown in Figure 13. The sensory flow enters the system from the environment, and encounters the top-down interpretation process of what we can refer to as the subjective mind. These two flows interact with each other intimately in between from which a new set of recognitions and actions is generated. The actions lead to changes in the environment which result in new sensations coming from the objective world. The recognition, on the other hand, causes the re-adaptation of the currently organized memory obtained through previous learning, which produces a new interpretation for the next sensation. What we see here is the open dynamic structure in which the relationship between the subjective mind and the objective world changes continuously through their mutual interaction.

Figure 13 is analogous to Varela's (Varela et al., 1991) views of the sense of the groundlessness of embodied cognition as well as to Matsuno's (Matsuno, 1989) view of the internal

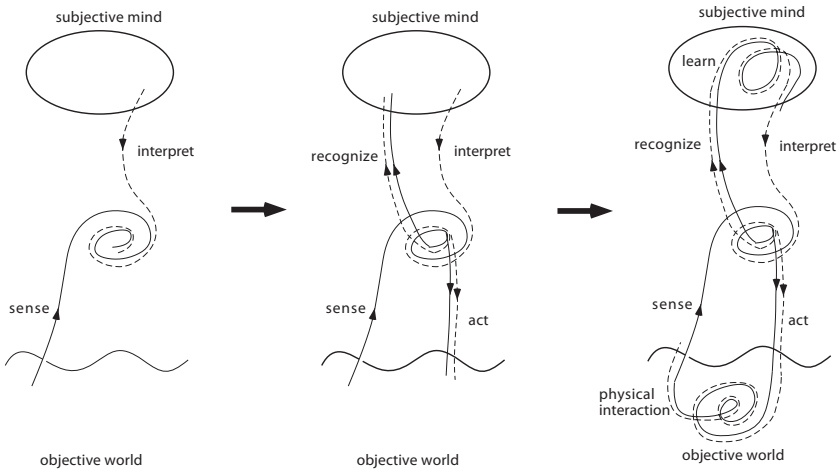

Figure 13: The open dynamic structure.

observer. Varela wrote that nothing can be found as responding to the ultimate existence of the subjective mind or the objective world since all one can find is their co-dependent structures. He argued that such structures can be grounded in neither side, but reside in between. Matsuno (Matsuno, 1989) and Gunji (Gunji & Konno, 1991) used the term "observation" as mostly equivalent to the term "interaction". They considered that the original relationship between the observer and the observed changes because of the interactions between them. This observer is called the "internal observer" since it is included in the internal loop of the interactions. The observation consists of a set of embodied processes that are physically constrained in various ways – e.g. delays in neural activation and body movement, limitation in learning capacity and so on. Such physical constraints in time and space do not allow the system to be uniquely optimized and given rise to incompleteness, nondeterminism and inconsistency. In fact in our robot experiment, such inconsistency arises in every aspect of system performances including recognition, learning and behaviour. However, at the moment of facing such an inconsistency, the processes cannot be terminated; instead, each process attempts to change its current relations to others in such a manner as if it were expected that the inconsistency will be resolved in the posterior time. It is interesting to note that the open dynamic structure maintains both stabilizing and unstabilizing mechanisms. The goal-directedness is an attempt to achieve the stability of the system by resolving the currently observed inconsistencies of the system. All processes of learning, perception, interpretation, behaviour and so on are regarded as goal-directed activities. However such goal-directed attempts always entail instability because of their embodiment as we have mentioned above. The co-existence of the stable and the unstable nature does not allow the system state to simply converge or diverge in its time-development. The trajectory of the system state is likely to show chaotic itinerary wandering among marginally stable attractors in a diverse manner as shown by Tsuda's (Tsuda, 1987) memory dynamics simulation using a chaotic neural network [3] .

[3]The essential difference of Tsuda's model from ours is that in Tsuda's formulation the instability originates from the non-equilibrium term explicitly represented in the neuro-dynamics model.

To the end of the discussion of the open dynamic structure, we would like to describe briefly its analogy to the Freeman analysis (Skarda & Freeman, 1987; Freeman, 1995) which is based on the electrophysiological studies of the olfactory systems of animals in the following two respects. First, in experiments on the conditioned behavior of an animal, a chaotic attractor appears in the neural activities in the olfactory systems when the animal experiences a new odorant. This means that the internal chaotic dynamics imply the novelty of the subjective experience. The second respect is that the correspondence between the spatio-temporal neural activation patterns and the odorants vary if the animal continues to learn new experiences day by day. This corresponds to our observation that the relation between the subjective mind and the objective world is always changing through incremental learning.

5.2 A model of the "structure of the self"

In this section, we first show our interpretation of the "self- consciousness" of the robot which has been obtained especially by observing the system dynamics characteristics of spontaneous phase transitions between the steady and the unsteady phases. Then, we discuss and evaluate this interpretation by considering possible correspondences in the literature on the discipline of phenomenology. To the end, we propose a possible model that represents the "structure of the self".

In the steady phase in our experiment, good coherence is achieved between the internal dynamics and the environmental dynamics when the subjective anticipation agrees well with observation. All the cognitive and behavioural processes proceed smoothly and automatically; no distinction can be made between the subjective mind and the objective world. In the unsteady phase, this distinction becomes rather explicit as the conflicts are generated between what the subjective mind expects and its outcome from the objective world. Consequently, we say that the "self-consciousness" of the robot arises in this moment of incoherence since the system's attention is now directed to the conflicts to be resolved. On the other hand, in the steady phase, the "self-consciousness" is diminished substantially since there are no conflicts to which the system's attention needs to be directed.

This interpretation of "self-consciousness" might be supported by Heidegger's (Heidegger, 1962) example of the hammer. For the carpenter, when everything is going smoothly, the carpenter himself and the hammer function as a unity. But, when something goes wrong with the carpenter's hammering or with the hammer, then the independent existences of the subject (the carpenter) and the object (the hammer) are noticed. Here, the carpenter becomes self-conscious just as he or she becomes conscious of the world as problematic. Another traditional phenomenologist, Merleau-Ponty (Merleau-Ponty, 1962), describes illness in the similar way. The healthy organism and environment function as a closely coupled unity in which we usually do not pay much attention to the organism. But when the organism becomes sick and goes wrong with the interaction with environment, then our attention goes to the organism. The essential claim shown in these two examples is quite analogous to our claim that "self-consciousness" emerges when the relation between the subject (top-down process) and the object (bottom-up process) becomes incoherent.

If we take it that "self-consciousness" and "non-self-consciousness" correspond to the unsteady and steady phases, respectively, then this leads to the further interesting idea that

the "self-conscious" situation takes place not as continuous in time but as discontinuous in time accompanied by the spontaneous phase transitions. This observation would correspond to Strawson's (Strawson, 1997) view in which he suggests the image of a string of pearls, as an image of a self. He claims that each self should be considered as a distinct existence, an individual thing or object, yet discontinuous as a function of time, as he inherited the idea from William James (1984). Hayward (Hayward, 1998) commented on Strawson's model by raising further questions: "do the continuous moments of experience themselves have structure; how are they held together so that we have the sense that they all belong to the same string." For these questions he investigates the structure of the self from the point of view of the Buddhist analysis of experience which is based on the disciplined examination of first-person experience through the method of shamatha-vipashyana meditation (Varela et al., 1991). From this study, he concludes that a particle-field analogy for the self is more apt than the string-of-pearls analogy. He claims that conscious experience is particle-like in that our sense of self appears to be ontologically distinct and relatively localized, but it also has a field-like and a non-local aspect.

Regarding this remark of Hayward, we further speculate that the field-like aspect is actually the dynamical systems aspect and also that the structure of the self corresponds to the dynamical structure of the system and not simply to what is captured by self-conscious operations. It is important to remember that all the characteristics of the time-development of the system are determined by this dynamical structure. Therefore, it is assumed that the particle-like aspect of self-consciousness accompanied by the phase transitions could be naturally explained when the structure of the self in terms of the dynamical structure of the system is well understood.

In order to clarify this idea, we first discuss the essential differences between machines and cognitive systems in general. A machine is basically operated in a way that its designer intended. Machines are designed to perform in a deterministic way such that their trajectory exactly repeats for the same sequences of inputs from the outside. In this sense, the performance of machines is completely controllable and observable from the outside. On the other hand, the behaviour of cognitive systems does not always follow this principle. For cognitive systems, it looks as if they maintained certain extents of autonomy to self-determine their own activities independent of the outside environmental interactions. Generally speaking, such systems encourage us to refer to their "selves". On the other hand, we cannot imagine any "selves" for machines, which are not allowed to maintain such autonomy. Then the question is how the autonomy of self-generating activities is enabled. Our answer is that such autonomy originates with the open dynamic structure which is characterized by the co-existence of the stable and the unstable mechanisms in terms of goal-directedness and embodiment respectively. As we have discussed previously, the open dynamic structure enables the system to continue to develop by changing the relation between the subjective mind and the objective world, accompanied by the spontaneous transitions between the unsteady and the steady phases. When the system exhibits a non-repeatable and diverse trajectory, then the "self" of the system is seen in the uniqueness of the trajectory. (On the other hand, if the trajectory converges into a stable one for good, no "self" would be seen.) Consequently, such a "self" becomes explicit internally (by observing the gaps between the subjective mind and the objective world), especially when the system comes across the unsteady phases in an unexpected manner. To the end of this section, our essential claims are summarized as:

1. There is essential structure of the "self" in the system and occurrences of "self-consciousness" are explained in terms of unfolding of this structure in time.

2. The structure of the "self" corresponds to the open dynamic structure which is characterized by stability in terms of goal-directed activities and instability caused by the embodiment.

3. When the system develops ever-changing relation between the subjective mind and the objective world in its non-repeatable trajectory, the uniqueness of the trajectory represents the "self" of the system.

4. The "self" becomes aware discontinuously when incoherence arises between the subjective mind and the objective world in a non-deterministic manner in the course of the time-development of the system.

6 Conclusion

We have attempted to explain our model of the "self" from our constructivist approach. Our experiments on robot learning showed that the incremental learning process evolves while the steady and unsteady phases appear intermittently. Our dynamical systems analysis showed that these fluctuations arise because of the intimate interaction between the bottom-up and the top-down processes. The comparison of this finding with phenomenological observation leads to a conclusion that the structure of the "self" corresponds to the dynamical structure of the system and that the "self" is made aware when the unsteady phase appears in the course of the time-development of the system.

What the current paper has shown is just one of many possible interpretations of the "self". It is expected that future collaborative studies between constructivists, phenomenologists and various empirical researchers would dramatically improve the current model of the "self".

References

[1] Amari, S. and Arbib, N. (1977) 'Competition and cooperation in neural nets', In J. Metzler, editor, *Systems Neuroscience*, pp. 119– 165. (San Diego: Academic Press)

[2] Ashby, W.R. (1952) *Design for a Brain.* (London, UK: Chapman and Hall).

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

[4] Blum, L., Shub, M. and Smale, S. (1989) 'On the theory of computational complexity over the real number', *Bulletine of the American Mathematical Society*, **21-1**, pp. 1–47.

[5] Braitenberg, V. (1984) *Vehicles: experiments in synthetic psychology.* (Cambridge, MA: MIT press).

[6] Brooks, R. (1991) 'Intelligence without representation', *Artif. Intell.*, **47**, pp. 139–159.

[7] Crutchfield, J.P. (1989) 'Inferring statistical complexity', *Phys Rev Lett*, **63**, pp. 105–108.

[8] Dennett, D. (1991) *Consciousness Explained.* (Boston, MA: Little Brown).

[9] Devaney, R.L. (1989) *An Introduction to Chaotic Dynamical Systems, Second Edition.* (Addison-Wesley).

[10] Elman, J.L. (1990) 'Finding structure in time', *Cognitive Science*, **14**, pp. 179–211.

[11] Freeman, W. (1995) *Societies of Brains: A Study in the Neuroscience of Love and Hate.* (Hillsdale, N.J: Erlbaum).

[12] Fuster, J.M. (1989) *The Prefrontal Cortex.* (New York: Raven Press).

[13] Gunji, Y.P. and Konno, N. (1991) 'Artificial Life with Autonomously Emerging Boundaries', *App. Math. Computation*, **43**, pp. 271–298.

[14] Harnad, S. (1990) 'The symbol grounding problem', *Physica D*, **42**, pp. 335–346.

[15] Hayward, J. (1998) 'A rDzogs-chen buddhist interpretation of the sense of self', *Journal of Consciousness Studies*, **5 (5-6)**.

[16] Heidegger, M. (1962) *Being and Time.* (New York: Harper and Row).

[17] Hopfield, J.J. (1986) 'Computing with neural circuits', *Science*, **233**, pp. 625–633.

[18] Hopfield J.J. and Tank, D.W. (1985) 'Neural computation of decision in optimization problems', *Biological Cybernetics*, **52**, pp. 141–152.

[19] Ikegami, T. and Taiji, M. (1998) 'Structure of possible worlds in a game of players with internal models', In *Proc. of the Third Int'l Conf. on Emergence*, pp. 601–604, Helsinki.

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

[21] James, W. (1984) 'Psychology: Briefer Course' (Cambridge, MA: Harvard University Press).

[22] Kawato, M., Maeda, Y., Uno, Y. and Suzuki, R. (1990) 'Trajectory formation of arm movement by cascade neural network model based on minimul torque-change criterion' *Biological Cybernetics*, **62**, pp. 275–288.

[23] Kuipers, B. (1987) 'A qualitative approach to robot exploration and map learning', In *AAAI Workshop Spatial Reasoning and Multi-Sensor Fusion*, pp. 774–779, Chicago.

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

[25] Matsuno, K. (1989) *Physical Basis of Biology.* (Boca Raton, FL: CRC Press).

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

[27] Merleau-Ponty, M. (1962) *Phenomenology of Perception* trans. Colin Smith (London: Routledge and Kegan Paul)

[28] Meyer, J.A. and Wilson, S.W. (ed. 1991) *From Animals to Animats: Proc. of the First International Conference on Simulation of Adaptive Behavior* (Cambridge, MA: MIT press).

[29] Moravec, H.P. (1982) 'The Stanford Cart and the CMU Rover', In *Proceeding of the IEEE*, **71-7** pp. 872–884.

[30] Nilsson, Nils J. (1984) 'Shakey the Robot' *SRI A.I. Center Technical Note 323.*

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

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

[33] Skarda C.A. and Freeman W.J. (1987) 'Does the brain make chaos in order to make sense of the world?', *Behavioral and Brain Sciences*, **10**, pp. 161–165.

[34] O'Keefe J. and Nadel L. (1978) *The hippocampus as a cognitive map,.* (Oxford, UK: Clarendon Press).

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

[36] Squire, L.R., Cohen, N.J. and 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).

[37] Strawson, G. (1997) 'The self', *Journal of Consciousness Studies*, **4 (5/6)**, pp. 405–428.

[38] Tani, J. and Fukumura, N. (1995) 'Embedding a grammatical description in deterministic chaos: an experiment in recurrent neural learning', *Biological Cybernetics*, **72**, pp. 365–370.

[39] Tani, J. (1996) 'Model-Based Learning for Mobile Robot Navigation from the Dynamical Systems Perspective', *IEEE Trans. System, Man and Cybernetics Part B, Special issue on learning autonomous robots*, **26-3**, pp.421–436.

[40] Tsuda, I. (1984) 'A hermeneutic process of the brain', *Progress of Theoretical Physics*, **79**, pp. 241–259.

[41] Tsuda, I., Koerner, E. and Shimizu, H. (1987) 'Memory dynamics in asynchronous neural networks', *Progress of Theoretical Physics*, **78**, pp. 51–71.

[42] Ungerleider L.G. and Mishkin. M. 'Two cortical visual systems', In D.G. Ingle, M.A. Goodale, and R.J. Mansfield, editors, *Analysis of Visual Behavior* (Cambridge, MA: MIT Press).

[43] van Gelder .T.J. (1998) 'The dynamical hypothesis in cognitive science', *Behavior and Brain Sciences* "in press".

[44] Varela, F.J., Thompson, E. and Rosch, E. (1991) *The embodied mind* (Cambridge, Mass: MIT Press).

[45] Walter, W.G. (1953) *The Living Brain.* (UK: Penguin).

[46] Waugh, F. and Westervelt, R. (1993) 'Analog neural nets with local competition, dynamics and stability', *Phys. Rev. E*, **47**, pp. 4524–4536.

[47] Wilson, M. (1994) 'Reactivation of hippocampal ensemble memories during sleep', *Science*, **265**, pp. 676–679.