Available online at www.sciencedirect.com
ScienceDirect
NEUROCOMPUTING
Neurocomputing 71 (2008) 3037–3043
www.elsevier.com/locate/neucom

Letters

Synaptic plasticity model of a spiking neural network for reinforcement learning

Kyoobin Lee[1], Dong-Soo Kwon*

Department of Mechanical Engineering, Korea Advanced Institute of Science and Technology (KAIST), ME 3042, 373-1 GuSung Dong YuSung Gu, Daejon 305-701, Republic of Korea

Received 19 March 2006; received in revised form 26 June 2007; accepted 9 September 2007
Communicated by R. Koetter
Available online 22 October 2007

Abstract

This paper presents a reward-related synaptic modification method of a spiking neuron model. The proposed algorithm determines which synapse is eligible for reinforcement by a reward signal. According to the proposed algorithm, a synapse is determined to be eligible when a presynaptic spike occurs shortly before a postsynaptic spike. A pre- and postsynaptic spike correlator (PPSC) is defined and used to determine synaptic eligibility, and to modify synaptic efficacy in cooperation with a reward signal. A simulation is conducted to demonstrate how the interaction between the PPSC and the reward signal influences synaptic plasticity.

Keywords: Spiking neural network; Synaptic plasticity; Reinforcement learning

1. Introduction

Reinforcement learning is useful for developing an intelligent agent without a teaching set. A reward signal during interaction with the environment can be thought as a source of learning. Several pioneering studies on animal behavior show the evidence of reinforcement learning in animals [8,19]. With the advancement in brain activity measurement, measuring the reinforcement learning activity through spiking patterns became possible. For example, we can measure the spiking pattern within a monkey's brain while he learns to push a button for a drop of sweat juice as a reward. In this experiment, the activity of dopamine neurons represents the reward signal (juice drop). This paper provides a biologically inspired algorithm that explains how the dopamine-like reward signal modifies the synaptic weights of a spiking neural network.

When animals learn in a reinforcement learning experiment, the learning sequence can be simplified as the following procedures:

1. A stimulus is given from environment.
2. An action is conducted by the neurons' spiking patterns generated by the stimulus.
3. The action changes environment.
4. The environmental change results in reward or punishment.
5. The given reward or punishment is converted to reward-related neuronal activity such as an elevation of dopamine concentration.
6. The reward/punishment signal modifies synaptic weights to increase/decrease the linkage between the stimulus and the selected action.

When we derive a synaptic reinforcement algorithm based on the above procedures, the algorithm has to consider two aspects: *eligibility* and *duration*. First, the algorithm should be able to determine which synapses are to be potentiated or depressed by the reward signal. In this paper, the term 'eligible synapse' is referred to the synapse that contributed to reward-earning. Second, when

*Corresponding author. Tel.: +82 42 869 3042; fax: +82 42 869 3210.
E-mail address: kwonds@kaist.ac.kr (D.-S. Kwon).
[1]Present address: Center for Neural Science, Korea Institute of Science and Technology, Seoul 136-791, Republic of Korea.

0925-2312/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2007.09.009

a synapse is determined as an eligible synapse, the eligibility has to sustain for a period of time. In general, a reward or punishment consequent upon an action does not come immediately. The temporal gap between an action and a reward or a punishment ranges from seconds to minutes. TD learning, a dynamic programming based reinforcement learning algorithm, uses 'eligibility trace' for this purpose.

The synaptic reinforcement has been studied by other researchers. Seung proposed a release-failure antagonism [14], in which the synaptic efficacy is potentiated by a reward signal when a presynaptic spike succeeds in releasing a postsynaptic spike. Conversely, when a presynaptic spike fails to release a postsynaptic spike, the synaptic efficacy is depressed. The premise of this paper is in general agreement with the basic idea of Seung's release-failure antagonism, and provides the explanations for different dynamics of the synaptic plasticity related to the reward signal. In the proposed algorithm, unlike the study of Seung, success or failure to release a postsynaptic spike is not determined at the time of every presynaptic spike, but determined by the dynamic model of intracellular substances. This dynamic model associates a postsynaptic spike with all presynaptic spikes prior to its time. Pfister et al. developed an optimal rule of synaptic change for STDP and related the result to reinforcement learning. In their model, the postsynaptic spike linked to a reward will be recreated by the synapse at the same time [9]. Their synaptic reinforcement model focuses on modifying the synaptic weight to make a postsynaptic spike occur at an exact timing to maximize the reward. However, our model focuses on finding synapses that are contributing to the reward and on modifying the eligible synapses by the reward signal.

2. Eligible synapse

The term 'eligible synapse' denotes a synapse that has contributed in obtaining a reward. Fig. 1 shows a situation in which signal flows in the brain cause a reward, such as dopamine secretion. For a series of neuronal activities, not all synapses of the neural network are contributing to the reward-earning; therefore, it is necessary to select only synapses that are eligible. In Fig. 1, the bold arrows denote the neuronal signal flows that may be considered for eligible synapse. Each number indicates the order of firing of neurons. The four dark areas, (a–d), represent four different cases of pre- and postsynaptic spike timing and can be described as:

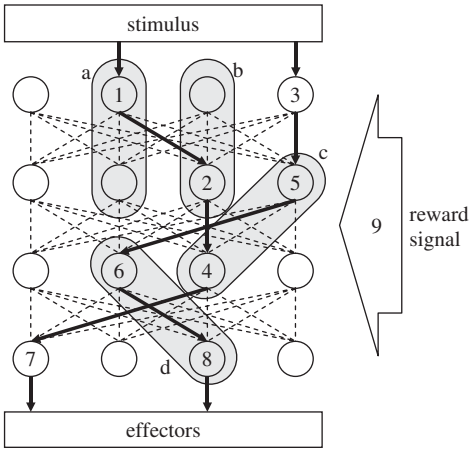


Fig. 1. Determination of eligible synapses. The number denotes the order of firing time.

(a) Only a presynaptic neuron fires without a postsynaptic spike (pre only).
(b) Only a postsynaptic neuron fires without a presynaptic spike (post only).
(c) A postsynaptic spike precedes a presynaptic spike (post-pre).
(d) A presynaptic spike precedes a postsynaptic spike (pre-post).

Synapses for cases (a) and (b) do not contribute to reward-earning and are not reinforced. For case (c), the synapse has to be weakened if the STDP rule is to be applied, but the synapse is left alone as it can be considered unrelated to reward-earning. The synapses satisfying case (d) are considered contributing to reward-earning and will be denoted as 'eligible synapse.' The determination of an eligible synapse is similar to the causal part of STDP, as they both have relevance to temporal causality.

Once a synapse is determined to be eligible for reinforcement, the synapse is eligible for rewards only in the near future. Such phenomenon can be justified in an example where an eligible synapse a year before should be thought as irrelevant to the current reward. Simply, a recent eligible synapse receives higher level of reinforcement compared to an older eligible synapse. It is reasonable to consider that the eligibility of a synapse monotonically decreases. Similarly, in TD learning, 'eligible trace' is defined as a mechanism for determining the magnitude of an update, and it decreases exponentially [17]. The proposed algorithm includes this feature and its detailed dynamics will be described in the following section.

3. Pre- and postsynaptic spike correlator (PPSC)

The temporal spike order is important in the determination of eligible synapses. As explained in Section 2, when a presynaptic spike precedes a postsynaptic spike, the synapse gains eligibility to be reinforced by future rewards, and then the synaptic eligibility decreases over time. A candidate for a dynamic equation that can determine eligible synapses is proposed as follows:

$$\frac{\mathrm{dPSI}(t)}{\mathrm{d}t} = \alpha(1-\mathrm{PSI}(t-\varepsilon))\delta_{\mathrm{pre}}(t) - \frac{\mathrm{PSI}(t-\varepsilon)}{\tau_{\mathrm{PSI}}}$$
$$\frac{\mathrm{dPPSC}(t)}{\mathrm{d}t} = \beta\mathrm{PSI}(t)(1-\mathrm{PPSC}(t-\varepsilon))\delta_{\mathrm{post}}(t) - \frac{\mathrm{PPSC}(t-\varepsilon)}{\tau_{\mathrm{PPSC}}} \quad (1)$$

where $\delta_{pre}(t) = \sum_{i=1}^{\infty}\delta(t-(i\text{th presynaptic spike time}))$, $\delta_{post}(t) = \sum_{i=1}^{\infty}\delta(t-(i\text{th postsynaptic spike time}))$, $\delta(t)$ is a Dirac delta function, α and β are constants, τ_{PSI} and τ_{PPSC} are time constants, and ε is a small time delay ($0<\varepsilon\ll 1$).

In order to determine eligible synapses, at least two intracellular substrates have to be introduced. One substrate, which increases when a presynaptic spike occurs, is termed a 'presynaptic spike indicator (PSI).' The other substrate, which increases when a presynaptic spike precedes a postsynaptic spike within a short period, is termed a 'pre- and postsynaptic spike correlator (PPSC).' Both the PSI and PPSC decrease exponentially and are bounded from 0 to 1. The PPSC not only determines an eligible synapse but also indicates the eligibility trace within a certain time period.

There may be more than two substrates for determining eligible synapses in real brains, and the dynamics are likely much more complicated and highly nonlinear. Using this simple model, several different patterns of pre- and postsynaptic pairs were tested as shown in Fig. 2: one presynaptic spike to one postsynaptic spike (a, e), one presynaptic spike to bursting postsynaptic spikes (b, f), bursting presynaptic spikes to one postsynaptic spike (c, g) and periodic spikes with a certain phase delay (d, h). In Fig. 2a–d, postsynaptic spikes precede presynaptic spikes. Therefore, the magnitude of PPSC is very small, indicating that the synapse is not eligible for reinforcement by the reward signal. In Fig. 2e–h, presynaptic spikes precede postsynaptic spikes, thus the value of PPSC shoots up and decreases exponentially. This indicates that the synapse is eligible for a certain period.

When a reward signal such as dopamine arrives while the PPSC has a positive quantity, the synaptic efficacy is modified by Eq. (2).

$$\frac{dw}{dt} = \eta R \times \text{PPSC} \quad \begin{cases} \text{if } w>1 \rightarrow w=1 \\ \text{if } w<0 \rightarrow w=0 \end{cases} \tag{2}$$

where R is the reward and $\eta>0$ is the learning rate.

The synaptic efficacy w is bounded between 0 and 1. The reward signal R represents the dopamine concentration around the synapse. Reynolds and Wickens provided a function that shows the relationship between dopamine concentration and synaptic change (See Fig. 4 in their paper) [10]. According to their function, a low dopamine concentration induces long term depression (LTD) and a high concentration induces long term potentiation (LTP). Therefore, it is assumed that R can be either positive or negative. This paper does not cover the modeling of dopamine concentration, which can be determined by the spike pattern of dopamine neurons, and R does not take the form of spike coding.

If R is assumed to be a positive constant, Eq. (2) is similar to the causal part of STDP. The main difference of the proposed reinforcement learning from STDP is that both LTP and LTD occur in a causal case (pre-before-post). Table 1 summarizes the relationship among spike timing, reward, STDP and reinforcement learning. In STDP, LTP occurs in a causal case and LTD occurs in an acausal case (post-before-pre). On the other hand, in reinforcement learning, both LTP and LTD occur in the causal case but the sign of the reward signal determines either LTP or LTD. In an acausal case, the proposed reinforcement learning does not change the synaptic efficacy because it is considered that the later presynaptic spike does not contribute to the sooner postsynaptic spike and the accompanying reward. Mathematically, this implies that the negative eligibility (PPSC<0) is not used.

Fig. 3 explains the mechanism of synaptic plasticity using the synaptic spike order, PSI, PPSC, and the reward signal. The reward signal is simplified as a rate-value rather than a spike-form though the dopamine concentration in the real brain is determined by the spike trains of dopamine

Table 1
Comparison between STDP and reinforcement learning

Spike timing	Sign of reward		
	$R>0$	$R=0$	$R<0$
STDP			
Pre-post	LTP	LTP	LTP
Post-pre	LTD	LTD	LTD
Reinforcement learning			
Pre-post	LTP	No change	LTD
Post-pre	No change	No change	No change

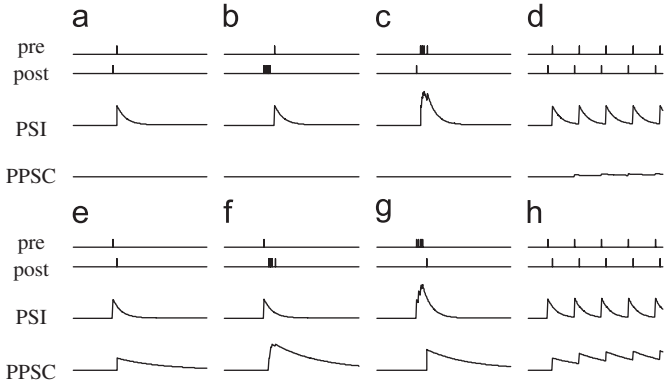


Fig. 2. A simple test of PSI and PPSC.

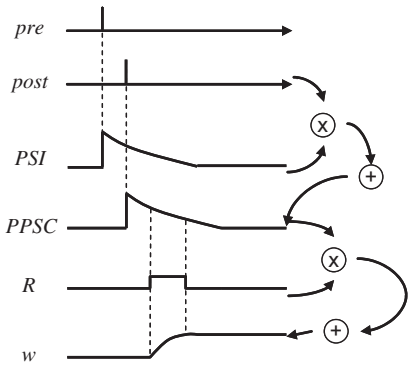


Fig. 3. Synaptic modification using PPSC and reward.

neurons. The dopamine system not only generates the reward signal but also learns the reward events and predicts them [4,5,12]. The modeling of a dopamine system compatible with the present algorithm is essential, but is not covered in this paper.

Although the proposed eligibility-determining algorithm is mathematically derived from the viewpoint of temporal causality, there is biological evidence supporting the algorithm. The calcium ion in synapses is believed to be related to the synaptic plasticity according to various experimental studies [2,3,11,16,21,22]. In these studies, a high calcium concentration causes LTP while a low calcium level causes LTD. The calcium concentration is determined by the synaptic spike orders. Senn et al. also developed a kinetic model of a NMDA receptor and calcium-activated secondary messenger for STDP [13]. The secondary messenger S_u in their model behaves similar to PPSC in the proposed algorithm but the time constant of PPSC is larger than the time constant of S_u because PPSC has to be associated with the reward that usually comes several seconds or minutes later. Although the aforementioned studies are related to the reward-independent synaptic plasticity such as STDP, there are several examples of biological evidence which suggest that the intracellular calcium interacts with the dopamine signal to contribute to reinforcement or memory [6,18]. Wickens and Kötter compiled ample studies to account for the dopamine reinforcement from various viewpoints for the requirements of plasticity, a calcium–dopamine relationship, an eligibility trace and a dopamine prediction [20]. For now, we do not specify the PPSC as the calcium level because other Ca^{2+}-activated substance such as calcium/calmodulin-dependent protein kinase II (CaMKII) can also be a candidate for the PPSC. Although the time constant of the intracellular calcium concentration is a nonlinear function of the concentration level, it can be roughly considered to be within 150 ms [15,16]. However, the time constant of calcium concentration is too short to be associated with the rewards because the rewards usually come after a few seconds or minutes later. Therefore, from the viewpoint of time constant, CaMKII seems to be a more reasonable candidate as it can sustain for a much longer period due to its autophosphorylation property [1,7]. After a more in-depth study, the PPSC could be shown to be the calcium ion, another intracellular substance activated by the calcium ion, or a completely different substance.

4. Simulation

In order to examine how the PPSC works in a spiking neural network, a simple simulation was conducted as follows (Fig. 4). A virtual robot flying over a $1\,m \times 1\,m$ workspace has a vision sensor of a 9×9 pixel array and four motors allowing movement with two degrees of freedom. The goal is located in the middle of the workspace and its location is detected by the 9×9 vision sensory array. The flying robot should try to position itself above the goal. Each motor is connected to one motor neuron. The motor neurons cause a small amount of motion in each direction (forward, backward, left, and right). All vision

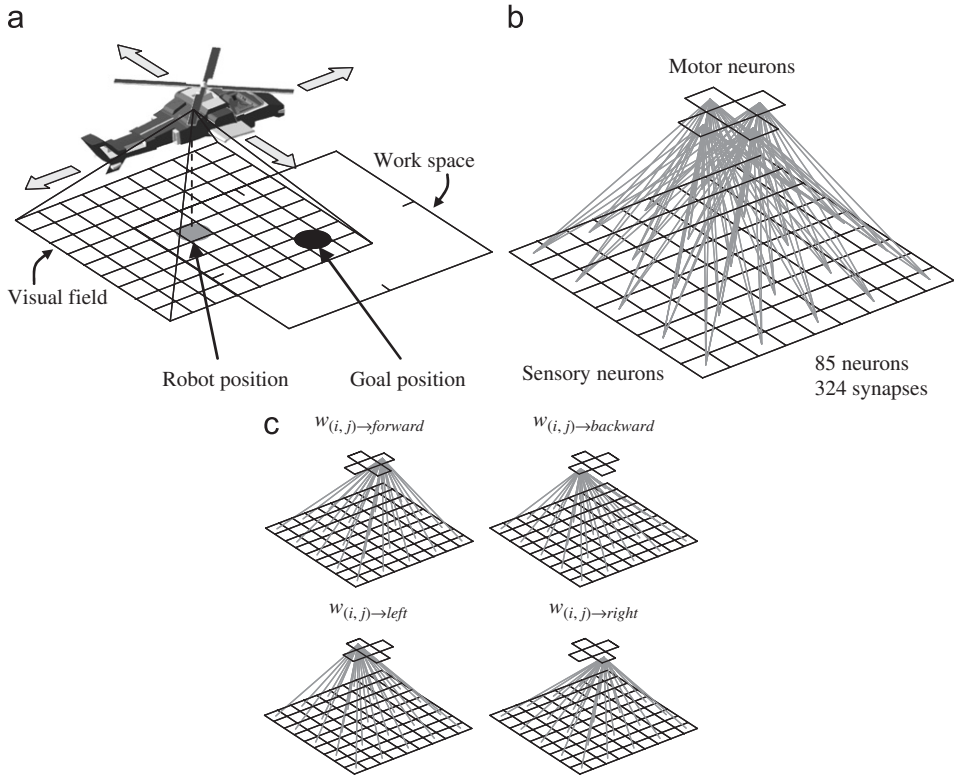


Fig. 4. Simulation setup.

sensor neurons are connected to the motor neurons (Fig. 4b) and the synaptic efficacies between the sensory neurons and the motor neurons are denoted by $w_{(i,j)\to\text{forward}}$, $w_{(i,j)\to\text{backward}}$, $w_{(i,j)\to\text{left}}$ and $w_{(i,j)\to\text{right}}$, where (i,j) indicates the position of the sensor pixel (Fig. 4c). The sensory neurons are attached to the robot so that they detect the relative position between the robot and the goal.

The dynamics of the sensory and motor neurons follows a linear integrate-and-fire neuron model:

$$\tau_m \frac{du(t)}{dt} = -u(t) + R_m I(t)$$
$$\text{if } u(t) \geqslant 1, \text{ then } u \leftarrow 0 : \text{ generate a spike} \quad (3)$$

where u is the nondimensional membrane potential, τ_m is the membrane time constant, I is the synaptic input current, and R_m is the membrane resistance.

The nondimensional membrane potential u is bounded between 0 and 1 as $u(t) = (v(t)-v_r)/(v_\vartheta - v_r)$ with the actual membrane potential $v(t)$, the threshold v_ϑ, and the rest potential v_r.

For the sensory neurons, a constant input current ($I = 2$) is given when the sensory neuron detects the goal. For the motor neurons, the input current is determined by the presynaptic spikes of the sensory neurons and the synaptic efficacies by Eq. (4).

$$I(t) = \sum_i \sum_j Q_{\max} w_i \frac{t}{\tau_s} \exp\left(-\frac{t}{\tau_s}\right) \Theta(t - t_{ij}) \quad (4)$$

where $Q_{\max}$ is the total charge when the synaptic efficacy w_i is 1, τ_s is the time constant for the input current, t_{ij} is jth firing time of the presynaptic neuron i, and $\Theta(t)$ is the Heaviside step function with $\Theta(t) = 1$ for $t \geqslant 0$ and $\Theta(t) = 0$ for $t < 0$.

There is one reward signal that dominates synaptic reinforcement. This reward signal can be either positive or negative. When the robot is at the goal, the reward signal is positive ($R = +\gamma$), while the signal is negative when the robot is out of the goal position ($R = -\gamma$). A positive or negative reward signal causes the eligible synapses to be potentiated or depressed. The initial values of synaptic weights are randomly chosen between 0.1 and 0.3. The initial values of u, PSI and PPSC are all 0. The simulation parameters are $\alpha = 0.1$, $\beta = 0.1$, $\tau_{PSI} = 10$ ms, $\tau_{PPSC} = 3$ s, $\eta = 1$, $\tau_m = 10$ ms, $R_m = 1$, $Q_{\max} = 0.02$, $\tau_s = 10$ ms, and $\gamma = 0.1$. The simulation time step is 1 ms. Pseudo-code of the learning stage is provided in Fig. 5.

Each spike of the motor neurons generates a small displacement (1 mm) of the robot in the corresponding direction. The direction and velocity of the movement are determined by the firing rates of the motor neurons. For example, if the firing rates of the motor neurons are 10, 32, 24, and 14 Hz for the directions of left, right, forward, and backward, respectively, then the direction and velocity of the robot movement will be as shown by the bold arrow in Fig. 6.

```
FOR ( ALL_SYNAPSE ) {                                   Initializes ynapses
        w = randombetween ( 0.1 , 0.3 ) }
WHILE ( SIMULATION_PERIOD ) {

ROBOT ( RANDOM_WALK )                                   Robot moves randomly

IF ( ROBOT_AT_GOAL)        R = +γ
ELSE IF (ROBOT_OUT_OF_GOAL)    R = -γ
                                                        Calculate reward
FOR (ALL_NEURONS) {
        u = u + (-u + R_m I) / τ_m · Δt
        IF ( u ≥ 1 )  u = 0 , GENERATE A SPIKE }
                                                        Calculate neuron dynamics
FOR ( ALL_SYNAPSE) {
        PSI = PSI - (PSI / τ_PSI) · Δt
        PPSC = PPSC - (PPSC / τ_PPSC) · Δt
        IF ( PRESYNAPTIC_SPIKE ) {
                PSI = PSI + α · (1 - PSI) }             Calculate eligibility
        IF ( POSTSYNAPTIC_SPIKE ) {
                PPSC = PPSC + β · PSI · (1 - PPSC) }

        w = w + η · R · PPSC · Δt
        IF ( w > 1 )  w = 1                             Update synaptic efficacy
        IF ( w < 0 )  w = 0  }
}
```

Fig. 5. Pseudo-code for simulation.

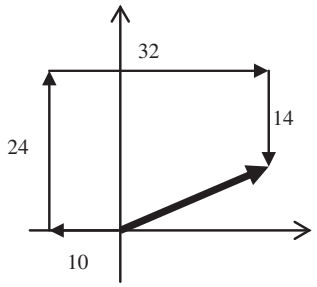


Fig. 6. An example for determining robot movement.

After learning for 1000 s, the behavior of the robot is illustrated in Fig. 7. The small circles denote the initial positions of the flying robot while the X-marks represent the final positions. The bold lines represent the velocity at which the robot moves. From every starting position, the robot approaches the goal position successfully. The time history of the synaptic efficacies is recorded in Fig. 8. There are four square patterns which denote the synaptic efficacies of the sensory-motor connections. The intensity of each pixel of the sub-square pattern indicates the

synaptic efficacies between the sensory neurons and motor neurons; a darker pixel indicates a higher synaptic efficacy. If the robot is positioned as shown in Fig. 4a, the goal is detected by the vision sensor at (2,8) and the sensor neuron (2,8) is activated consequently. This activation influences the motor neurons with the synaptic efficacies of $w_{(2,8)\to\text{forward}}$, $w_{(2,8)\to\text{backward}}$, $w_{(2,8)\to\text{left}}$, and $w_{(2,8)\to\text{right}}$ and is represented by four dots at 200 s in Fig. 8. As time passes, the synaptic efficacies gradually turn into a distinctive pattern such that the robot achieves a behavior to move toward the goal. As a result, the synaptic efficacies for forward and right movements ($w_{(2,8)\to\text{forward}}$ and $w_{(2,8)\to\text{right}}$) are strengthened as depicted by darker shade, while the synaptic efficacies for left and backward movements ($w_{(2,8)\to\text{left}}$ and $w_{(2,8)\to\text{backward}}$) are weakened as depicted by much lighter shade.

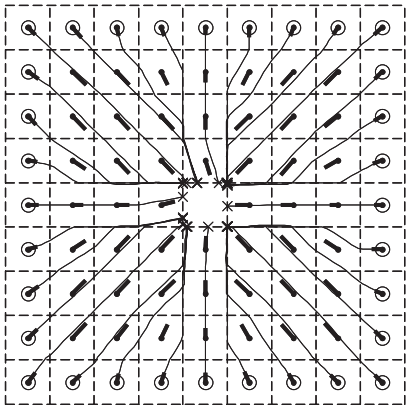


Fig. 7. The movement of the robot after 1000 s of learning (drawn in the global coordinate of the workspace).

5. Conclusions and future work

The use of the PPSC is proposed for reinforcement learning in a spiking neural network. The PPSC is used to determine the synaptic pathway eligible for reward, and each synapse was reinforced only if the postsynaptic spike occurred shortly after a presynaptic spike. The magnitude of synaptic update exponentially decreases as a function of time due to the property of the PPSC. The proposed method was evaluated through a simulation with 85 neurons and 324 synapses for a goal-finding task based on the input from the 9×9 pixel array of a vision sensor. Training the neurons for 1000 s resulted in the propensity of the neurons to drive the mobile unit of the robot to the goal.

Some may argue that the biologically inspired approach is difficult to justify due to its inherent difficult nature of performing experiment. But this should not confine us within the intellectual domain where every phenomenon has been already verified with experiments. Our history revealed that synthesizing a model which cannot be verified at the time of its establishment should not be ruled out. When the Hebbian learning was first introduced from Hebb's postulate, he did not prove its validity through biological experiments at that time. But now, we all accept its validity. We believe that both the computational neuroscientists and the biological neuroscientists should look at each other's approaches to complement the need in explaining how nature takes its course.

In regard to the contribution of the present research, we must say it is the method of determining the eligible synapses using the dynamic models of two intracellular substances. This introduction of intercellular substances is expected to provide inspiration to biological experiments.

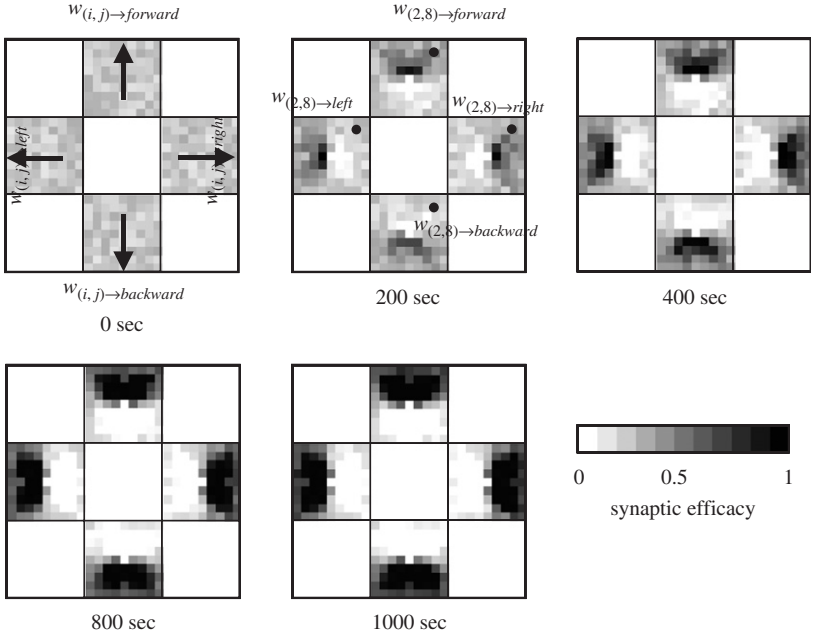


Fig. 8. Transition of synaptic efficacies during learning.

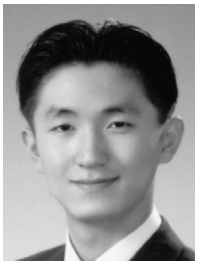
Understanding the relationship between the parameters (α, β, τ_{PSI}, and τ_{PPSC}) in the PSI and PPSC and the rate of environment change is critical for learning efficiency. A fast environmental change would require smaller values of the time constants, but the optimal parameters with respect to different rates of environmental change have yet to be determined. The modeling of a dopamine system compatible with a spiking neural network is also required for reward prediction. In an animal test that examines the dopamine prediction, it has been shown that the animals can learn from an experimental environment and make decisions better due to dopamine prediction and reinforcement learning [12]. In a certain reinforcement case, an earlier action should be encouraged rather than a later action. For example, in the game of 'noughts and crosses,' the first move is more important than the last move. This example might mislead us into believing that the eligibility of the first move should be larger than that of the last move. However, the reward prediction property of dopamine neurons can explain this. The predicted reward causes less dopamine release than the unpredicted reward. Therefore, if the reward is predicted at the time of performing the last move by repetitive learning, the synaptic change for the last move by dopamine reinforcement becomes weak. The STDP, reinforcement learning and dopamine prediction model will be integrated and tested using a mobile robot in a real-world environment.

Acknowledgments

This research was performed for the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs funded by the Ministry of Commerce, Industry and Energy of Korea. This work was also supported by the Korea Research Foundation. (Grant no. KRF-2003-041-D00026).

References

[1] U.S. Bhalla, R. Iyengar, Emergent properties of networks of biological signaling pathways, Science 283 (1999) 381–387.
[2] R.J. Cormier, A.C. Greenwood, J.A. Conner, Bidirectional synaptic plasticity correlated with the magnitude of dendritic calcium transients above a threshold, J. Neurophysiol. 85 (2001) 399–406.
[3] J.A. Cummings, R.M. Mulkey, R.A. Nicoll, R.C. Malenka, Ca^{2+} signaling requirements for long-term depression in the hippocampus, Neuron 16 (1996) 825–833.
[4] N.D. Daw, D.S. Touretzky, Long-term reward prediction in TD models of the dopamine system, Neural Comput. 14 (2002) 2567–2583.
[5] K. Doya, Metalearning and neuromodulation, Neural Netw. 15 (2002) 495–506.
[6] T.M. Jay, Dopamine: a potential substrate for synaptic plasticity and memory mechanisms, Prog. Neurobiol. 69 (2003) 375–390.
[7] J. Lisman, H. Schulman, H. Cline, The molecular basis of CaMKII function in synaptic and behavioural memory, Nat. Rev. Neurosci. 3 (2002) 175–190.
[8] J. Olds, P. Milner, Positive reinforcement produced by electrical stimulation of the septal area and other regions of the rat brain, J. Comp. Physiol. Psychol. 47 (1954) 419–427.
[9] J.-P. Pfister, T. Toyoizumi, D. Barber, W. Gerstner, Optimal spike-timing-dependent plasticity for precise action potential firing in supervised learning, Neural Comput. 18 (2006) 1318–1348.
[10] J.N.J. Reynolds, J.R. Wickens, Dopamine-dependent plasticity of corticostriatal synapses, Neural Netw. 15 (2002) 507–521.
[11] J.E. Rubin, R.C. Gerkin, G.-Q. Bi, C.C. Chow, Calcium time course as a signal for spike-timing-dependent plasticity, J. Neurophysiol. 93 (2005) 2600–2613.
[12] W. Schultz, Getting formal with dopamine and reward, Neuron 36 (2002) 241–263.
[13] W. Senn, H. Markram, M. Tsodyks, An algorithm for modifying neurotransmitter release probability based on pre- and postsynaptic spike timing, Neural Comput. 13 (2000) 35–67.
[14] H.S. Seung, Learning in spiking neural networks by reinforcement of stochastic synaptic transmission, Neuron 40 (2003) 1063–1073.
[15] H.Z. Shouval, M.F. Bear, L.N. Cooper, A unified model of NMDA receptor-dependent bidirectional synaptic plasticity, Proc. Natl. Acad. Sci. 99 (2002) 10831–10836.
[16] H.Z. Shouval, G.C. Castellani, B.S. Blais, L.C. Yeung, L.N. Cooper, Converging evidence for a simplified biophysical model of synaptic plasticity, Biol. Cybern. 87 (2002) 383–391.
[17] R.S. Sutton, A.G. Barto, Reinforcement Learning: An Introduction, MIT Press, Cambridge, MA, 1998.
[18] T. Suzuki, M. Miura, K.-Y. Nishimura, T. Aosaki, Dopamine-dependent synaptic plasticity in the striatal cholinergic interneurons, J. Neurosci. 21 (2001) 6492–6501.
[19] E.L. Thorndike, Animal intelligence: an experimental study of the associative processes in animals, Psychol. Rev. Monogr. Suppl. 2 (4) (1898).
[20] J. Wickens, R. Kötter, Cellular models of reinforcement, in: J.C. Houk, J.L. Davis, D.G. Beiser (Eds.), Models of Information Processing in the Basal Ganglia, The MIT Press, Cambridge, MA, 1998.
[21] S.-N. Yang, Y.-G. Tang, R.S. Zucker, Selective induction of LTP and LTD by postsynaptic $[Ca^{2+}]i$ elevation, J. Neurophysiol. 81 (1999) 781–787.
[22] L.C. Yeung, B.S. Blais, L.N. Cooper, H.Z. Shouval, Calcium as the associative biochemical signal for a model of Hebbian plasticity, Neurocomputing 52–54 (2003) 437–440.



Kyoobin Lee received the B.S. degree in mechanical engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 1998, and the M.S. degree in mechanical engineering from KAIST 2000. He is currently working towards the Ph.D. Degree in Human-Robot Interaction Research Center in KAIST. His research interests include brain-inspired robot intelligence, human-robot interaction, medical robotics, haptics and neuroscience.

Dong-Soo Kwon received the B.S degree in mechanical engineering from the Seoul National University, Korea in 1980, the M.S. degree in mechanical engineering from Korea Advanced institute of Science and Technology (KAIST), Seoul, Korea in 1982 and the Ph.D. degree in mechanical engineering from Georgia Institute of Technology, Atlanta, Georgia, USA, in 1991. From 1991 to 1995, he was a research staff at Oak Ridge National Laboratory. He is currently a Professor of mechanical engineering and the director of Human-Robot Interaction Research Center at KAIST, Daejon, Korea. His current research interests include human-robot/computer interaction, telerobotics, medical robots, and haptics. He is a member of IEEE, KSME and ICASE.