Scientific
Research

Motor Learning Based on the Cooperation of Cerebellum and Basal Ganglia for a Self-Balancing Two-Wheeled Robot

Xiaogang Ruan, Jing Chen, Lizhen Dai
School of Electronic Information and Control Engineering, Beijing University of Technology, Beijing, China
E-mail: chenjing0828@139.com

Abstract

A novel motor learning method is present based on the cooperation of the cerebellum and basal ganglia for the behavior learning of agent. The motor learning method derives from the principle of CNS and operant learning mechanism and it depends on the interactions between the basal ganglia and cerebellum. The whole learning system is composed of evaluation mechanism, action selection mechanism, tropism mechanism. The learning signals come from not only the Inferior Olive but also the Substantia Nigra in the beginning. The speed of learning is increased as well as the failure time is reduced with the cerebellum as a supervisor. Convergence can be guaranteed in the sense of entropy. With the proposed motor learning method, a motor learning system for the self-balancing two-wheeled robot has been built using the RBF neural networks as the actor and evaluation function approximator. The simulation experiments showed that the proposed motor learning system achieved a better learning effect, so the motor learning based on the coordination of cerebellum and basal ganglia is effective.

1. Introduction

Many skills and behaviors of human or animals are formed and developed gradually in the self-learning and self-organization process of nervous system. It is important to understand and imitate the internal learning mechanism of human and animals' nervous system, and it is an important issue to give such a mechanism to machine for control science, artificial intelligence and robotics study. Opposite with passive machinery reception learning, autonomous learning is an active self-knowledge and independent learning, and it is relatively the basis of agent's long-term learning and development. The interdisciplinary across of psychology, neurophysiology and machine learning theory promoted the production and development of autonomous learning theory and method directly.

Related researches of neurophysiological theories indicate that, during the sensorimotor learning, the cerebral cortex, basal ganglia, and cerebellum work in parallel and unique way [1-3]. The ballpark behavior adapt to the environment can be learnt through the basal ganglia loop

by reinforcement learning [4], and the ballpark behavior can be refined by supervised learning in the cerebellum [5]. Cerebral cortex, utilizing unsupervised Hebbian learning [6-7], driven by input from basal ganglia and cerebellum, learns through practice to perform these operations fast and accurately. Studies show that the basal ganglia are the medium of behavior selection, and nucleus in basal ganglia, such as striatum, globus pallidus, substantia nigra, *et al.*, play the role of action se- lection [8-11]. Bernard W. Balleine, et al, from psychology and brain research institute, university of California, has studied the integrative function of basal ganglia in operant learning [12], which indicates that the dorsal striatum in basal ganglia plays an important role in the operant conditioning learning process. From above introductions, it can be seen that human brain is a complex system, and the basal ganglia and cerebellum in the central nervous system all play an important role in motor learning. For motor learning, two different mechanisms can operate to make use of sensory information to correct motor errors. The fist is on-line error correction, and the second is trial-to-trial learning, in the latter case, errors from one

trial influence the motor output on the next trial. The two learning mechanisms are related with the cerebellum and basal ganglia.

However, the motor control and learning mechanism of cerebellum and basal ganglia seems be studied separately. For example, the cerebellum works as an adaptive filter [13,14] and the feedback error learning in cerebellum is proposed by Kawato [15,16]. An important class of BG model is known as Actor-Critic models or RL-based models [17,18] with system-level. In 2008, a contracting basal ganglia model was proposed for the action selection mechanism [19]. Although Doya [20] pointed the complementary roles of cerebellum and basal ganglia, how to implement such cooperation is not very specific. This paper is dedicated to the coordinated learning mechanism between cerebellum and basal ganglia.

In the aspect of learning control theory, professor Wang proposed a kind of hybrid algorithm aiming at the obstacle avoidance problem of robot [21]. They proposed a neural fuzzy system with mixed coarse learning and fine learning phases. In the first phase, supervised learning method is used to determine the membership functions for the input and output variables simultaneously. After sufficient training, fine learning is applied which employs reinforcement learning algorithm to fine tune the membership. Meng Joo Er *et al.* also applied hybrid learning approach for the obstacle avoidance [22]. In the simulation environment, preliminary supervised learning was executed applying dynamic fuzzy neural net- works, and in the real robot, neuro-fuzzy controller is capable of re-adapting in a new environment. After carrying out the learning phase on a simulated robot, the controller is implemented on a real robot. The hybrid learning method shows the coordination between supervised learning and reinforcement learning.

In the aspect of learning algorithm research, Michael *et al.* combined supervised learning with actor-critic learning and proposed a supervised learning combined with an actor-critic architecture [23]. This type of learning has been suggested by Clouse and Benbrahim [24,25]. In this paper, the supervised actor-critic algorithm is associated with the interactions between the cerebellum and basal ganglia. At the same time, operant learning will be considered in the design of motor learning system as the action selection mechanism in the basal ganglia. The example of motor learning in a self-balancing robot will be present to show the benefits of using the proposed motor learning method.

There are many kinds of two-wheeled self-balancing robot, such as wheeled-inverted pendulum, self-balancing wheelchair, JOE, nBot [26-29], and so on. All of them have ability of keeping balanced themselves and running forward. For the balancing control, researchers have done a lot of works for such kinds a robot. An adaptive output recurrent cerebellar model articulation controller is utilized to control wheeled inverted pendulums (WIPs) [30]. Grasser *et al.* [28] developed a dynamic model for designing a mobile inverted pendulum controller using a Newtonian approach and linearization method. In [26], a dynamic model of a WIP was created with wheel motor torques as input, accounting for nonholonomic no-slip constraints. Vehicle pitch and position were stabilized using two controllers. Jung and Kim [31] created a mobile inverted pendulum using neural network (NN) control combined with a proportional-integral-derivative controller. But there are fewer reports for the autonomous learning based on the neuro-physiological theory. Balancing skill's acquisition is the results of central nervous system. How to imitate the brain function and construct a motor learning method for balancing robot is the main target in this paper.

This paper is organized as follows. Section 2 addresses the motor learning in the central nervous system of human being and shows the important role of the cerebellum and basal ganglia in CNS. Section 3 introduces operant learning method. That is an important working style in the motor learning and it is closely related with the mechanism of the basal ganglia. Section 4 is the main part of this paper. Motor learning method with the interaction of the cerebellum and basal ganglia is proposed. Evaluation mechanism, action selection mechanism and tropism mechanism are included. The convergence analysis is deduced in the sense of entropy. At last, in section 5, the control structure is designed with the proposed method for a self-balancing two-wheeled robot and the comparative experiments are conducted to show the effectiveness of proposed method. Section 6 concludes.

2. Motor Learning in CNS

An important feature of the central nervous system (CNS) in its control of movement is the capability of motor learning. Motor learning is the process of improving the motor skills, the smoothness and accuracy of movements. It is important for calibrating simple movements like reflexes, as parameters of the body and environment change over time. As we know, for higher mammals, especially humans, both of supervised learning and reinforcement learning are probably the most important class of motor learning. In the CNS, two auxiliary monitor systems are the cerebellum and ganglion nuclear group in the forebrain area. Cerebral cortex, cerebellum and basal ganglia, cooperated with each other, form complex motor control system. **Figure 1** shows their relations in CNS [32]. Cerebellum specialize in supervised learning
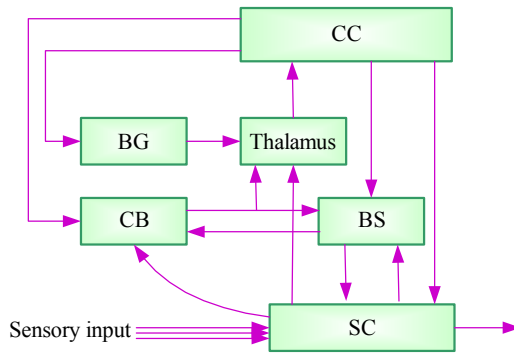
Figure 1. Schematic of relationship for central nervous system (CNS). Cerebral Cortex (CC); Basal Ganglia (BG); Cerebellum (CB); brainstem (BS); spinal cord (SC). In the CNS, BG and CB play an important role in motor control and learning. The thalamus relays sensation, spatial sense and motor signals to the cerebral cortex.

and its training signals derive from the climbing fiber input from the inferior olive, as shown in **Figure 2**. Basal ganglia are involved in reinforcement learning with the operant conditioning role and the learning signals are dopaminergic signals from Substantia Nigra. The neural circuit of basal ganglia is shown in **Figure 3**.

3. Operant Learning

3.1. The Related Works about Operant Learning

In psychology, operant is a class of behavior that produces consequences by operating (*i.e.*, acting) upon the environment. Operant conditioning (OC) is a technique of behavior modification through reinforcement and punishment. The research about operant conditioning theory [34] was started in 1938 by Skinner, a psychology professor. Its consequence influences the occurrence and form of behavior. Operant conditioning and classical conditioning [35] are two main learning ways of associative learning, and all animals, including human, have these two learning way. Operant conditioning is distinguished from classical conditioning in that operant conditioning deals with the modification of operant behavior. Operant conditioning reflects the relation between behavior and its outcome, and the learning with OC theory is called operant learning (*i.e.* instrument learning). Recently, researchers apply OC theory in the robot learning and control and have done plenty of experimental studies. For example, Björn Brembs *et al.* [36] from Germany applies himself to the research of the operant conditioning in flies (Drosophila) and snails. 'Pure' operant conditioning and parallel operant conditioning at the flight simulator were studied. Chengwei, Yao *et al.* [37] applied OC theory into emotion development and presented an emotion development agent model based on OCC
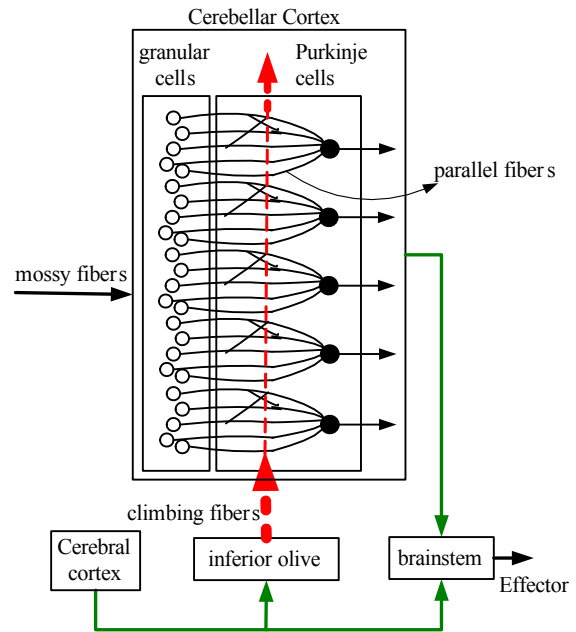

Figure 2. Working principle of cerebellar cortex. Climbing fibers carry the error signals to modify the PF-PC (Parallel fibers-Purkinje cells) synapse through LTD (long-term depression) [33].
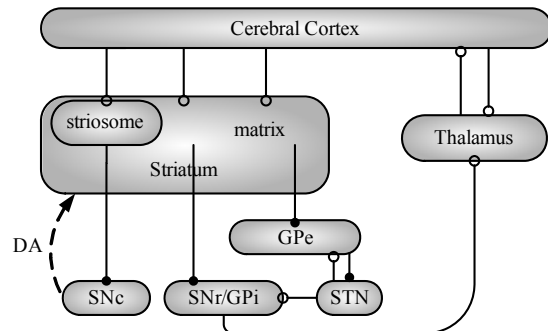

Figure 3. Neural circuit of the basal ganglia. Substantia nigra pars compacta (SNc); substantia nigra pars reticulate (SNr); internal segment of globus pallidus (GPi); external segment of globus pallidus (GPe); subthalamic nucleu (STN); dopamine (DA); ○, excitatory connection; ●, inhibitory connection. DA from substantia nigra is feedback to striatum as the learning signals.

Model and operant conditioning. In, Kazuko Itoh *et al.* [38] in Japan built a behavior model of humanoid robots based on operant conditioning, where the robot could select and output its behavior according to its behavior potential and evaluation of the behavior. Robot WE-4RII learned the behavior of shaking hand with people using his right hand. Bernard *et al.* [12] had studied the integrative function of BG with operant conditioning, which suggested that the dorsal striatum in BG play an important part in operant learning. That is why the OC theory

is considered in the research of BG mechanism. The principle of OC is present in the following section.

3.2. The Principle of Operant Conditioning Theory

Operant conditioning emphasizes that behavior's consequence influences the occurrence. The difference between operant learning and supervised learning depends on the feedback signal obtained from environment, which is valuation signal or error signal. There are three main elements about the learning control based on OC theory, including probabilistic action selection mechanism, evaluation mechanism and tropism mechanism, which is the main part of learning and is used to update the action selection policy. **Figure 4** shows the learning control mechanism based on OC theory. The probabilistic action selection mechanism is the main part of OC learning.

4. CB-BG-Based Motor Learning Scheme

4.1. CB-BG-Based Motor Learning Control Structure

According to the working modes of the cerebellum and basal ganglia, a CB-BG-based motor learning scheme is presented, which exploits supervised ac- tor-critic learning structure, as shown in **Figure 5**.

The coordinated factor ω is used to calculate the combined action a, which can be got through the following formula.

$$a \leftarrow \omega a_E + (1-\omega)a_S \qquad (1)$$

Thereinto, a_E, the exploratory action, can be got by an artificial neural networks addition to a probabilistic action selection mechanism. a_S, the supervised action, can be got by the traditional feedback controller, which provides intention and solution for the action networks. a, a simple weighted sum of the actions given by the component policies, is sent to the environment.

4.2. Algorithm Description

1) *Evaluation Networks with BG mechanism*: The evaluation function $V(s)$ can be updated with TD error[39]: $\delta(t) = r(t) + \gamma V(s_{t+1}) - V(s_t)$. z_t is the eligibil- ity trace of evaluation networks weights, which is com- puted by the linear differential equation, defined as,

$$z_t = \gamma \lambda z_{t-1} + \frac{\partial V(s_t)}{\partial W_t} \qquad (2)$$

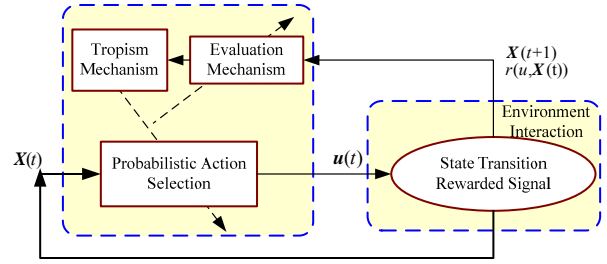The updating equation of evaluation networks based

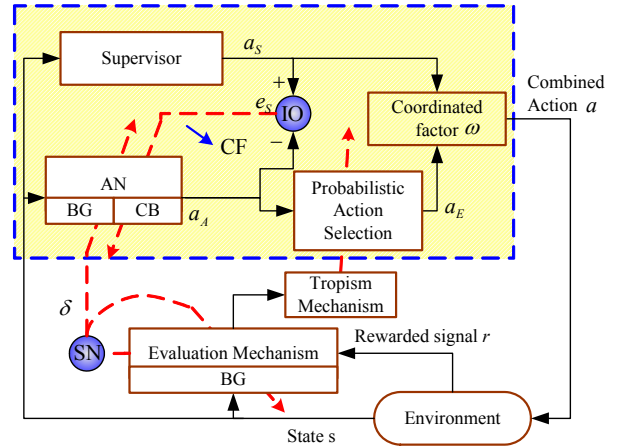**Figure 4. Learning control mechanism based on OC theory.**

Figure 5. CB-BG-based motor learning control structure. BG: Basal Ganglia; CB: Cerebellum; IO: Inferior Olive; SN: Substantia Nigra; CF: Climbing Fibre; AN: Action Net-works, which is constituted with the interaction of BG and CB; Evaluation Mechanism owns to BG. The real line is the data flow, while the dashed line part is the learning updating algorithm. Dopaminergic signal (δ) coming from SN and feedback to striatum in BG is the learning signal for the action network in BG part. Another learning signal is the supervised error e_S coming from the Inferior Olive (IO) in correspondence with the learning mechanism of the cerebellum. The two parts of learning signals can be coordinated by coordinated factor ω Probabilistic Action Selection (PAS) mechanism expresses the operant learning in BG. a_E, the exploratory action; a_S, the supervised action; r represent the reward signal after executing the combined action; state s represent the input of action networks and evaluation networks.

on eligibility trace is that

$$w(t) \leftarrow w(t) + \beta \delta(t) z_t \qquad (3)$$

$\beta > 0$ is a positive constant and express the learning step of evaluation networks; $0 < \gamma \leq 1$ is the discount factor of reward; $0 \leq \lambda < 1$ is the decay factor of eligibility trace. If $\delta(t) > 0$ (*i.e.*, the state is better than expect), the estimated critic value $V(s)$ would be increased. In the contrast, if $\delta(t) < 0$ (*i.e.*, it is worse than expect), the estimated value $V(s)$ in the state would be decreased.

2) *Action Networks with the interaction of CB and BG*: In **Figure 5**, the actor policy $\pi_A(s)$ is a mapping from state to actor and can be given by a function approximator with the parameter vector θ.

$$a_A = \pi_A(s, \theta) \tag{4}$$

The random action selective strategy a_E obeys a probabilistic distribution of Boltzmann-Gibbs [40].

$$\varphi(a_E) = \frac{1}{Z} \exp\left\{-\frac{\varepsilon(a_E)}{K_B T}\right\} \tag{5}$$

Thereinto, $\varepsilon(a_E) = (a_E - a_A)^2, T > 0$ is the thermodynamics temperature, $K_B = 1.38 \times 10^{-23}$ Js/K is the Boltzmann constant, $\exp\{-\varepsilon(a_E)/K_B T\}$ is Boltzmann factor, Z is the partition function, and that:

$$Z = \int_{-\infty}^{+\infty} \exp\left\{-\frac{\varepsilon(a_E)}{K_B T}\right\} da_E = \sqrt{K_B T \pi} \tag{6}$$

After each state transition, those parameters are modified according to the following rule. It includes two parts, $\Delta\theta^{CB}$ is realized through the function of cerebellum, and $\Delta\theta^{BG}$ is correspondence with the basal ganglia part.

$$\theta \leftarrow \theta + \omega\Delta\theta^{BG} + (1-\omega)\Delta\theta^{CB} \tag{7}$$

In order to compute the supervised learning update: $\Delta\theta^{CB}$, the supervisory error can be defined for each observed state as follows.

$$E(s) = \frac{1}{2}\left[\pi_S(s) - \pi_A(s)\right]^2 \tag{8}$$

The weights updating of the action networks in cerebellum part is that:

$$\Delta\theta^{CB} = -\alpha\nabla_\theta E(s) = \alpha(a_S - a_A)\nabla_\theta\pi_A(s) \tag{9}$$

The actor networks updating in BG part applies a kind of approximated policy gradient estimation algorithm:

$$\Delta\theta^{BG} = \alpha\delta(a_E - a_A)\nabla_\theta\pi_A(s) \tag{10}$$

$\alpha > 0$ is the learning step of action networks; $0 \le \omega \le 1$ weighs between two kinds of gradient information, one is related with second evaluation signal δ (also called internal return) and another gradient information is related with the supervised error $(a_S - a_A)$. In order to express the importance of rewarded signal, internal return signal is introduced into the item $\Delta\theta^{CB} = \alpha\delta(a_S - a_A)\nabla_\theta\pi_A(s)$, so we get the improved weights updating method.

$$\theta \leftarrow \theta + \alpha\delta\left[\omega(a_E - a_S) + (a_S - a_A)\right]\nabla_\theta\pi_A(s) \tag{11}$$

The improved supervised actor networks weight-updating algorithm shows that the TD error modulates the supervisory error $(a_S - a_A)$. That is to say, one 'trusts'

the critic more than the supervisor, in which case one should view the supervisor as an additional source of exploration. That is the coordination between the cerebellum and basal ganglia. When the internal return δ is small, the actor network loses the ability of learning through the supervisor. While in the unimproved algorithm, this learning ability only depends on the change of coordinated factor ω.

3) *Tropism Mechanism*: In [41], the temperature parameter defined by the state values and the frequency of visiting states are added to representative state-action. By this thought, the probabilistic tropism can be changed through the temperature T, which depends on the critic value $V(s)$. The bigger is the estimated value $V(s)$, the better the action tropism is, and the temperature is more smaller, and vice versa. The smaller $V(s)$, the worse tropism, the higher temperature, and system is in much more stochastic situation. The tropism rule is:

$$T = \frac{k_1}{1 + \exp(k_2 V(s_t))} \tag{12}$$

Thereinto, $k_1 > 0, k_2 > 0$, are constant and k_1 is defined as **Simulated Annealing factor(SAF)**, which can be changed with learning; $V(s_t)$ is the critic value of current state.

4) *Interaction Mechanism of the cerebellum and basal ganglia*: The coordinated factor ω expresses the weight of cerebellum supervisory learning in the whole learning process. It can be changeable and reflect the role of cerebellum and basal ganglia in different motor learning stages. The following exponential increase method can be used as in literature [42] shown in **Figure 6**.

$$\omega = \max\left\{0, 1 - \exp\left(-\frac{t - n_2}{n_3}\right)\right\}(n_2, n_3 > 0) \in C \tag{13}$$

5) *CB-BG-based Motor Learning Algorithm with the RBF Networks*: A radial basis function (RBF) network, which is an artificial neural network that uses radial basis functions as activation functions, is proposed by J. Moody and C. Darken in the late 1980s. It is a linear combination of radial basis functions. RBF networks imitate the neural networks structure of local adjustment and mutual covered receptive fields in human brain. The
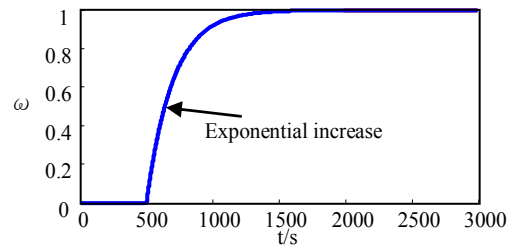

Figure 6. CB coordination factor.

radial basis function can be used for the coding of CMAC. The conventional CMAC uses constant binary or triangular receptive-field basis function. However, the derivative information is not preserved. To acquire derivative information of input and output variable, Chiang and Lin developed a CMAC with a differentiable Gaussian receptive-field basis function and provided convergence analysis results for this network [43]. As mentioned above, we will approximate action and evaluation functions by RBF neural networks in our proposed algorithm.

The three layers topological structure of RBF networks is shown in **Figure 7**, which is introduced in detail as follows.

- The first is the input layer, which is for input space. Each neuron is corresponding to an input variable, and given that $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)^{\mathrm{T}} \in \Re^n$.

- The second is the hidden layer, in which every node is a Gaussian function. The vector formed by output of hidden layer, $\boldsymbol{\phi} = (\phi_1, \phi_2, \cdots, \phi_{n_{\mathrm{B}}})^{\mathrm{T}}$, n_{B} is called the resolution. The bigger n_{B} has a higher resolution.

$$\phi_i = \exp\left[-\sum_{k=1}^{n} \frac{(x_k - c_{ki})^2}{2b_{ki}^2}\right] \qquad (14)$$

where, $\boldsymbol{c}_i = [c_{1i}, c_{2i}, \cdots, c_{ni}]^{\mathrm{T}} \in \Re^n$ is the center vector; $\boldsymbol{b}_i = [b_{1i}, b_{2i}, \cdots, b_{ni}]^{\mathrm{T}} \in \Re^n, (i = 1, 2, \cdots, n_{\mathrm{B}})$ is the base width vector.

- The third is the output layer, which have one node, used for approximating the unknown nonlinear function. The weights vector is: $\boldsymbol{v} = [v_1, v_2, \cdots, v_{n_{\mathrm{B}}}]^{\mathrm{T}}$, and the output is got.

$$y_{\text{out}} = \boldsymbol{v}^T \boldsymbol{\phi} = \sum_{i=1}^{n_{\mathrm{B}}} v_i \phi_i \qquad (15)$$

For the action and evaluation networks, the same structure of RBF neural networks is used. The output weights vector of actor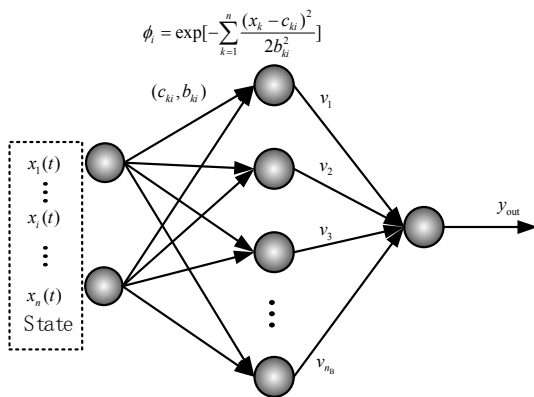 and evaluation networks are represented with $\boldsymbol{\theta}$ and $\boldsymbol{w}$ respectively. So the process of CB-BG-based motor learning algorithm with the RBF networks can be concluded as follows:

CB-BG-based Motor Learning Algorithm:

Step 1: Initialization: Initiate the learning step α, β; discount factor $\gamma \in (0,1]$; eligibility trace delay factor $\lambda \in [0,1)$; Initiate arbitrarily all weights parameters $\boldsymbol{\theta}$ and $\boldsymbol{w}$; Let *trytime* $= 0$.

Step 2: Repeat (for each episode):

1) $t \leftarrow 0$；$z_0 \leftarrow 0$ (reinitialize eligibility trace), Choose an initial state s_0.

2) Repeat (for each step in the episode):

 a) For current state s_t, calculate $a_A \leftarrow \pi_A(s_t)$; $a_S \leftarrow \pi_S(s_t)$; Select exploration action a_E according to Boltzmann distribution; set the coordinated factor ω; compute the combined action a and the eligibility trace $z_t = \gamma\lambda z_{t-1} + \boldsymbol{\phi}(s_t)$.

 b) Execute action a, receive immediate reward $r(t)$, then observe the new state s_{t+1}; calculate TD error $\delta(t) = r(t) + \gamma V(s_{t+1}) - (s_t)$; update the weights through $\boldsymbol{w}_{(t+1)} \leftarrow \boldsymbol{w}_{(t)} + \beta\delta(t)z_t$; until satisfied the condition $\angle(\boldsymbol{w}(t+1), \boldsymbol{w}(t)) < \varepsilon$; $\boldsymbol{\theta}_{(t+1)} \leftarrow \boldsymbol{\theta}_{(t)} + \alpha\delta\left[\omega(a_E - a_S) + (a_S - a_A)\right]\boldsymbol{\phi}(s_t)$

 c) $t \leftarrow t+1$;

Until the desired number of episodes has been investigated.

4.2. Algorithm Convergence Analyses

Theorem 1: Action a_E obeys probability distribution of Boltzmann-Gibbs: $\varphi(a_E) = \dfrac{1}{Z}\exp\left\{-\dfrac{\varepsilon(a_E)}{K_B T}\right\}$, where,

$\varepsilon(a_E) = (a_E - a_A)^2$,

$Z = \int_{-\infty}^{+\infty} \exp\left\{-\dfrac{\varepsilon(a_E)}{K_B T}\right\} \mathrm{d}a_E = \sqrt{K_B T \pi}$, The sum of the probability in real sets is one.

Proof: The sum of the probability in real sets can be compute in the following integral equation.

$$\sum_p = \int_{-\infty}^{+\infty} \varphi(a_E) \mathrm{d}a_E = \int_{-\infty}^{+\infty} \frac{1}{Z}\exp\left\{-\frac{\varepsilon(a_E)}{K_B T}\right\} \mathrm{d}a_E$$

$$= \frac{1}{\sqrt{K_B T \pi}} \int_{-\infty}^{+\infty} \exp\left\{-\frac{\varepsilon(a_E)}{K_B T}\right\} \mathrm{d}a_E = \frac{\sqrt{K_B T \pi}}{\sqrt{K_B T \pi}} = 1$$

Theorem 2: Along with learning, the entropy of system: $E = -\int_{-\infty}^{+\infty} \varphi(a_E)\log_2(\varphi(a_E)) \mathrm{d}a_E$ decreases gradually, and system changes from uncertainty to certainty as time.

Proof: The sum of the probability in real sets can be compute in the following integral equation.

Figure 7. Architecture of a RBFN.

Literature [44] listed several entropy of continuous probability distribution, and the entropy of continuous system with obeys normal distribution is:

$E = \frac{1}{2}\log_2\left(2\pi e\sigma^2\right)$. It can be got that the probability distribution of Boltzmann-Gibbs in the paper is consistent with normal distribution: $N\left(a_s, \sqrt{\frac{K_BT}{2}}\right)$, that is to say, the mean is $\mu = a_s$ and the variance

$\sigma = \frac{Z}{\sqrt{2\pi}} = \sqrt{\frac{K_BT}{2}}$.

So that, the entropy of the continuous system is:

$$E = \frac{1}{2}\log_2\left(\pi e K_BT\right) \tag{16}$$

Along with the learning progress, $V(s)$ approximates the largest sum of discounted reward. According to the tropism mechanism as formula (12), and by setting suitable parameters, the explore and exploit can keep coordinated, and the action policy which has been learnt at last can satisfy $\pi_E^* = \arg\max_{\pi_E} V_{\pi_E}(s)$. The learning algorithm makes the function $V(s)$ developing toward the maximum direction. From the tropism mechanism, we can find that T would decrease with the increase of V. The system transforms from uncertainty to certainty along with learning.

Entropy of discrete system has better characters, such as, non-negative characteristic and coordinates transformation invariability, etc. Different with discrete system, entropy of continuous system, also called cross-entropy or differential entropy, does not have characteristics above. When $T < 1/\left(\pi e K_B\right) = 8.4855\times10^{21}\left(\text{K}\right)$, $E < 0$, the entropy could be negative as shown in **Figure 8**. Therefore, the proposed method is convergence in the sense of entropy, and the system changes from uncertainty to certainty.

5. Simulation Experiment and Analyses

5.1. Self-Balancing Two-Wheeled Robot

Self-balancing two-wheeled robot is a complex nonlinear under-actuated system and has self-balancing skill like human. There is important theory signification to research about the cognition mode of robot. The schematic structure is shown in **Figure 9**.

Parameters are defined in **Table 1** which is shown in Appendix.

The dynamic mathematic model is

$$M\left(\tilde{q}\right)\ddot{\tilde{q}} + N\left(\tilde{q}, \dot{\tilde{q}}\right) = ET \tag{17}$$

where, $\tilde{q} = \left(\theta_b, \varphi_l, \varphi_r\right)^T$, $T = \begin{bmatrix} \tau_l & \tau_r \end{bmatrix}^T \in \Re^{2\times1}$, the ma-
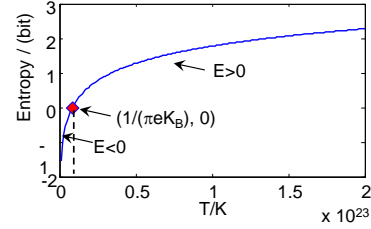

Figure 8. Entropy changing with temperature T.
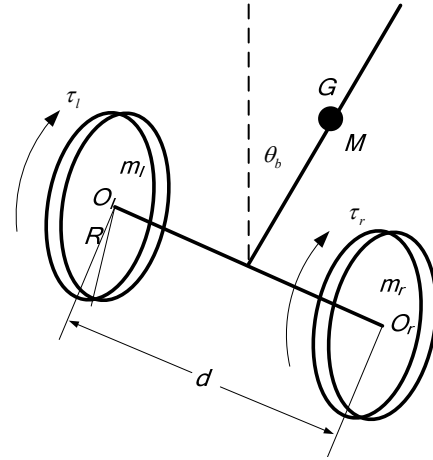
Figure 9. System structure of robot.

Table 1. Definition of Parameters.

Parameter	Description	Value
m_l, m_r /kg	mass of the left and the right wheel	1
M/kg	the mass of the intermediate body	10
R/m	radius of wheel	0.15
d/m	distance between the 0_l and 0_r	0.44
L/m	distance between the 0 and G	0.4
$J_l = J_r = J_w$/kg·m²	the moment of inertia of (left and right wheel) about its axis	0.01125
J_y/kg·m²	the moment of inertia of the robot about the y –axis	1.5417
J_z/kg·m²	the moment of inertia about the z –axis	0.5893
g/m/s²	acceleration of gravity	9.8
μ_l, μ_r /(N·m/(rad/s))	frictional coefficient of two wheels with ground	0.1,0.1
μ_w /(N·m/(rad/s))	frictional coefficient of the wheels axis	0.1
x/m	forward displacement of robot	variable
v/(m/s)	forward velocity of robot	variable
θ_b /rad	inclination angle of the intermediate body	variable
$\dot{\theta}_b$ / rad/s	inclination angle velocity of the intermediate body	variable
θ/rad, $\dot{\theta}$/rad/s	yaw angle and angle velocity of robot	variable
φ_r, φ_l /rad	rotary angle of the right and left wheel	variable
ω_l, ω_r / rad/s	rotary angle velocity of the right and left wheel	variable
τ_l /N.m	torque provided by left motor	[–5, 5]
τ_r /N.m	torque provided by right motor	[–5, 5]

trix is illustrated in appendix.

According to the definition of under-actuated system, the affined nonlinear system can be formulated with $\ddot{\tilde{q}} = f_1\left(\tilde{q},\dot{\tilde{q}},t\right) + f_2\left(\tilde{q},\dot{\tilde{q}},t\right)u$, and if the following condition is satisfied: $rank\left[f_2\left(\tilde{q},\dot{\tilde{q}},t\right)\right] < \dim[\tilde{q}]$ (*i.e.* the number of actuators is less than DOF, degree of freedom), the system is an under-actuated system. As for the two-wheeled self-balancing robot, $rank\left[M^{-1}\left(\tilde{q}\right)E\right] = 2 < 3$, so that it is an under-actuated system. Equation (16) is left multiplied transformation matrix A, $AM\left(\tilde{q}\right)\ddot{\tilde{q}} + AN\left(\tilde{q},\dot{\tilde{q}}\right) = AET$, where,

$A = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$. Then we can get:

$$\Psi\left(\tilde{q}\right)\ddot{\tilde{q}} + \Phi\left(\tilde{q},\dot{\tilde{q}}\right) = E'T \tag{18}$$

where, $E' = AE = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}^{\mathrm{T}}$

Equation (18) can be divided into:

$$\begin{cases} \Psi_{11}\ddot{\tilde{q}}_1 + \Psi_{12}\ddot{\tilde{q}}_2 + \Omega_1\left(\tilde{q},\dot{\tilde{q}}\right) = 0 \\ \Psi_{21}\ddot{\tilde{q}}_1 + \Psi_{22}\ddot{\tilde{q}}_2 + \Omega_2\left(\tilde{q},\dot{\tilde{q}}\right) = T \end{cases} \tag{19}$$

where, $\tilde{q} \in \Re^n, \tilde{q}_1 \in \Re^l, \tilde{q}_2 \in \Re^m, l = n - m$, $\tilde{q}_1$ denote all the passive joints, and $\tilde{q}_2$ denote the actuated joints; Ω denotes the Coriolis force and gravity terms.

$\Psi\left(\tilde{q}\right) = \begin{bmatrix} \Psi_{11} & \Psi_{12} \\ \Psi_{21} & \Psi_{22} \end{bmatrix}$, where Ψ is positive definite, Ψ_{11} and Ψ_{22} are also positive definite. Through the controllability analysis, we can conclude that the system is controllable. So it is important that under-actuated system is not necessarily an uncontrollable system.

5.2. Simulation Experiment

Applying the motor learning method proposed in Section IV, we did a simulation research on a self-balancing two-wheeled robot for the balance skill learning, and the bionic learning control structure is illustrated in **Figure 10**. There are mainly five parts in the control system: receptor, afferent nerve, nerve centre, efferent nerve and effector. The proposed CB-BG-based motor learning method in Section 4 is included in the nerve centre. DAS denotes data acquisition system.

According to the sliding mode control thought, we define the rewarded function and the sliding surface H as follows:

$$H : \left\{x \mid s(x) = 0\right\} \tag{20}$$

$$s(t) = a_1\theta_b(t) + a_2\dot{\theta}_b(t) \tag{21}$$

We can obtain the reward information according to the changing of the sliding surface. After executing an action, if $s(t)$ changes near to zero, give it reward. Conversely, give it penalty. 0 denotes reward, and -1 denotes penalty. It can be formulated as:

$$r_t = \begin{cases} 0 \text{ (reward)} & \text{if} \quad s^2(t+1) \le s^2(t) \\ -1 \text{ (penalty)} & \text{else} \end{cases} \tag{22}$$

Detect the robot states signals, and normalize the input signals which are sent to the CNS including the cerebellum and basal ganglia. The coordinated factor ω is used in the form of exponential decay as formula (13). The parameters are chosen as follows. T_s is the sampling time. $n_{\theta_b}, n_{\dot{\theta}_b}$ are the normalizing factor for angle and angle velocity separately.

$T_s = 10ms;\ n = 2;\ n_B = 49;\ k_1 = 10;\ k_2 = 0.1;\ \alpha = 0.1;$

$\beta = 0.1;\ \lambda = 0.8;\ \gamma = 0.8;\ n_2 = 500;\ n_3 = 200;\ a_1 = 10;$

$a_2 = 1;\ n_{\theta_b} = \pi/54;\ n_{\dot{\theta}_b} = \pi/54;$

$c_1 = c_2 = [-3, -2, -1, 0, 1, 2, 3]^{\mathrm{T}};$

$b_1 = b_2 = [0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5]^{\mathrm{T}}$

$w_a = \boldsymbol{\theta} \in R^{49\times 1};\ w_c = \boldsymbol{w} \in R^{49\times 1};\ \tau_l = \tau_l = a;$

$a_S = 50\theta_b + 10\dot{\theta}_b + \left(\varphi_l + \varphi_l\right)/2 + \left(\varphi_l + \varphi_l\right)/2$

The initial inclination angle $\theta_b(0) = 0.1744(rad)$.

1) *The first experiment:* two experiments are done for contrast. One with cerebellum coordinated, and another without cerebellum coordinated. Observe the result of inclination angle θ_b.
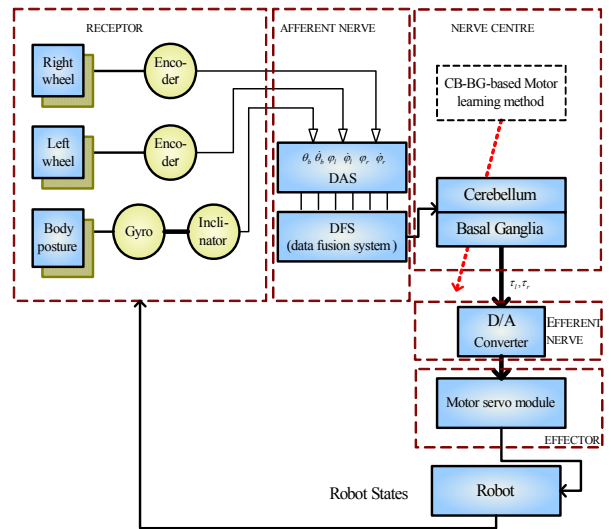

Figure 10. The bionic motor learning structure for self-balancing two-wheeled robot.

From the results, we can conclude that robot can learn to choose suitable action intelligently and keep balance with the two methods. The different is that, with the co-ordination of cerebellum, the robot can achieve the expected result with 0 failure times, shown in **Figure 11(a)**. However, the learning method with only basal ganglia mechanism, the robot has failed for 8 times and got the correct action in the ninth try time, and the Figure and data analysis chart are shown in **Figures 11(b)** and **(c)**.

2) *The Second Experiment:* After learning with cerebellum coordinated for one try time, robot can keep balance through operant learning only. Adjust the Simulated Annealing factor (SAF) k_1, and observe the inclination angle changing.

The process of annealing is as follows: In the beginning, the SAF: $k_1 = 10$; on the 11th turn learning: $k_1 = 10 \rightarrow k_1 = 5$; on the 16th turn learning: $k_1 = 5 \rightarrow k_1 = 1$; on the 21st turn learning: $k_1 = 1 \rightarrow k_1 = 0.5$; on the 23rd turn learning: $k_1 = 0.5 \rightarrow k_1 = 0.1$.

The average temperature and entropy for every turn have been calculated as shown in **Figure 12**. It can be found that the entropy is decreasing with learning. Observe the balancing learning results and the inclination angle curves are illustrated in **Figure 13**, and the convergence results have been contrasted in the beginning and the end in **Figure 14**. From the experiments results, the convergence can be seen with the decreasing of temperature.

From above two experiments, it can be shown that, using the proposed CB-BG-based motor learning method, the self-balancing robot can learn the balancing skill high efficiently. Through the decreasing the temperature, learning results have become better and better. And the decreasing of entropy also indicates that the system is developing from uncertainty to certainty.

6. Conclusions

For the behavior learning of agent, a novel motor learning method has been present. That is based on the cooperation of the cerebellum and basal ganglia. The evaluation mechanism, tropism mechanism, action selection mechanism and the learning algorithm with the cerebellum coordination mechanism have been proposed. In the beginning, the learning signals come from two parts (Inferior Olive and Substantia Nigra). At last, the learning signals only come from Substantia Nigra. Through analysis, the algorithm is convergent in the sense of entropy. With a self-balancing two-wheeled robot as the experiment plant, a motor learning control system has been constructed for the robot with the proposed method, in which RBF neural network is used to approximate the
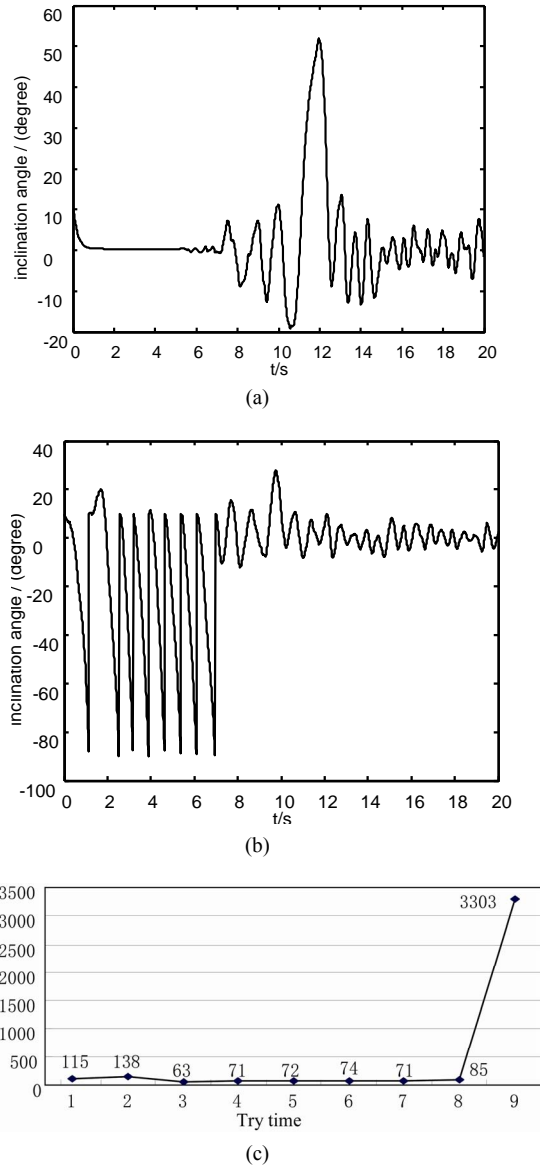

(a)

(b)

(c)

Figure 11. Results of comparative experiments. (a) Simulation result with the proposed motor learning method; (b) Simulation result with the traditional actor-critic method; (c) Data analysis with unimproved method.
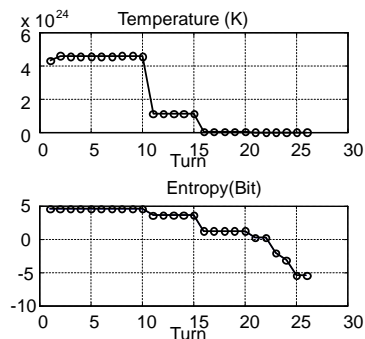

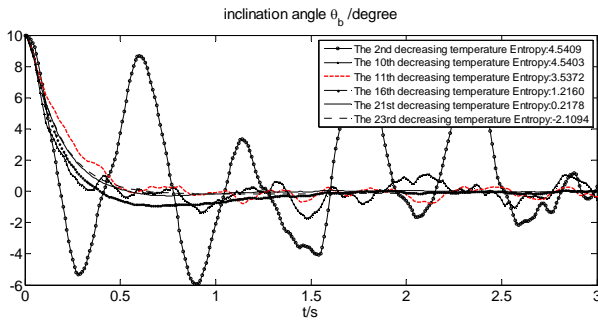**Figure 12. Temperature and entropy changing with learning.**

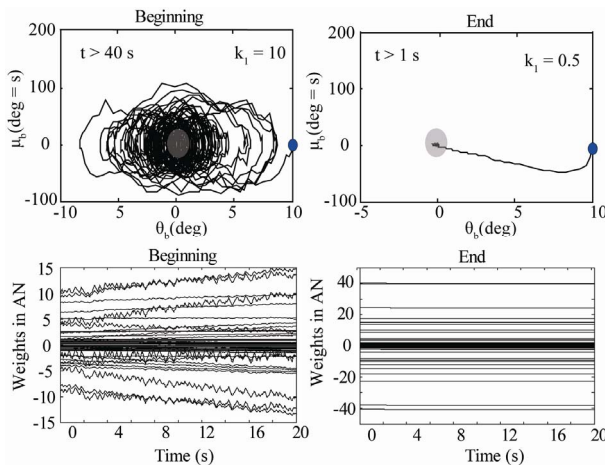**Figure 13. Changing of inclination angle with the decreasing of temperature.**

Figure 14. Convergence process of learning.

action and evaluation networks, use the sliding mode method to design the reward function, use the exponential decay factor to design the cerebellar effect in the whole motor learning process. The contrast of simulation experiments shows that the proposed method has improved the learning algorithm with only basal ganglia mechanism in the aspect of learning speed and failure time. After many rounds of learning and the temperature decreasing, the certainty of system has been reinforced eventually along with the decreasing of entropy. The vibration is disappearing eventually, and the learning speed and accuracy have further improved. The principal contributions of this paper are that a motor learning method on the neurophysiology with the coordination of cerebellum and basal ganglia, on which a central nervous system for a self-balancing two-wheeled robot can be constructed.

7. Acknowledgements

8. References

[1] J. C. Houk and S. P. Wise, "Distributed Modular Architectures Linking Basal Ganglia, Cerebellum, and Cerebral Cortex: Their Role in Planning and Controlling Action," *Cerebral Cortex*, Vol. 5, No. 2, 1995, pp. 95-110. doi: 10.1093/cercor/5.2.95

[2] X. Lu, O. Hikosaka and S. Miyachi, "Role of Monkey Cerebellar Nuclei in Skill for Sequential Movement," *Journal of Neurophysiology*, Vol. 79, No. 5, 1998, pp. 2245-2254.

[3] K. Doya, "What are the Computations of the Cerebellum, the Basal Ganglia and the Cerebral Cortex?" *Neural Networks*, Vol. 12, No. 7-8, 1999, pp. 961-974. doi:10.1016/S0893-6080(99)00046-5

[4] J. C. Houk, "Agents of the mind," *Biological Cybernetics*, Vol. 92, No. 6, 2005, pp. 427-437. doi:10.1007/s00422-005-0569-8

[5] E. I. Knudsen, "Supervised Learning in the Brain," *The Journal of Neuroscience*, Vol. 14, No. 7, 1994. pp. 3985-3997.

[6] T. V. P. Bliss and G. L. Collingridge, "A Synaptic Model of Memory: Long-Term Potentiation in the Hippo-Campus," *Nature*, Vol. 32, No. 2, 1993, pp. 31-39. doi:10.1038/361031a0

[7] S. E. Hua and J. C. Houk, "Cerebellar Guidance of Premotor Network Development and Sensorimotor Learning," *Learning and Memory*, Vol. 4, No. 1, 1997, pp. 63-76. doi:10.1101/lm.4.1.63

[8] B. Girard, N. Tabareau, Q. C. Phama, A Berthoz and J. J. Slotine, "Where Neuroscience and Dynamic System Theory Meet Autonomous Robotics: A Contracting Basal Ganglia Model for Action Selection," *Neural Networks*, Vol. 21, No. 4, 2008, pp. 628-641. doi:10.1016/j.neunet.2008.03.009

[9] R. Bogacz and K. Gurney. "The Basal Ganglia and Cortex Implement Optimal Decision Making between Alternative Actions," *Neural Computation*, Vol. 19, No. 2, 2007, pp. 442-477. doi:10.1162/neco.2007.19.2.442

[10] W. M. Jonathan, "The Basal Ganglia: Focused Selection and Inhibition of Competing Motor Programs," *Progress in Neurobiology*, Vol. 50, No. 4, 1996, pp. 381-425. doi:10.1016/S0301-0082(96)00042-1

[11] M. X. Cohen and M. J. Frank, "Neurocomputational Models of Basal Ganglia Function in Learning, Memory and Choice," *Behavioural Brain Research*, Vol. 199, No. 1, 2009, pp. 141-156. doi:10.1016/j.bbr.2008.09.029

[12] B. W. Balleine, M. Liljeholm and S. B. Ostlund, "The Integrative Function of the Basal Ganglia in Instrumental Conditioning," *Behavioural Brain Research*, Vol. 199, No. 1, 2009, pp. 43-52. doi:10.1016/j.bbr.2008.10.034

[13] P. Dean and J. Porrill, "Adaptive-Filter Models of the

 ICA

Cerebellum: Computational Analysis," *The Cerebellum*, Vol. 7, No. 4, 2008, pp. 567-571. doi:10.1007/s12311-008-0067-3

[14] P. Dean, J. Porrill, C. F. Ekerot E and J. Henrik, "The Cerebellar Microcircuit as an Adaptive Filter: Experimental and Computational Evidence," *Nature Reviews Neuroscience*, Vol. 11, No. 1, 2010, pp. 30-43. doi:10.1038/nrn2756

[15] M. Kawato, "Feedback-Error-Learning Neural Network for Supervised Motor Learning," Elsevier, Amsterdam, 1990.

[16] M. Kawato and H. Gomi, "A Computational Model of Four Regions of the Cerebellum Based on Feeback-Error-Learning," *Biological Cybernetics*, Vol. 68, No, 2, 1992, pp. 95-103. doi:10.1007/BF00201431

[17] D. Joel, Y. Niv and E. Ruppin, "Actor-Critic Models of the Basal Ganglia: New Anatomical and Computational Perspectives," *Neural Network*, Vol. 15, No. 4-6, 2002, pp. 535-547. doi:10.1016/S0893-6080(02)00047-3

[18] M. Khamassi, L. Lachèze, B. Girard, A. Berthoz and A. Guillot, "Actor-Critic Models of Reinforcement Learning in the Basal Ganglia: from Natural to Artificial Rats," *Adaptive Behavior*, Vol. 13, No. 2, 2005, pp. 131-148. doi:10.1177/105971230501300205

[19] B. Girard, N. Tabareau, Q. C. Phama, A. Berthoz and J. J. Slotine, "Where Neuroscience and Dynamic System Theory Meet Autonomous Robotics: A Contracting Basal Ganglia Model for Action Selection," *Neural Networks*, Vol. 21, No. 4, 2008, pp. 628-641. doi:10.1016/j.neunet.2008.03.009

[20] K. Doya, "Complementary Roles of Basal Ganglia and Cerebellum in Learning and Motor Control," *Current Opinion in Neurobiology*, Vol. 10, No. 6, 2000, pp. 732-739. doi:10.1016/S0959-4388(00)00153-7

[21] C. Ye, N. H. C. Yung and D. W. Wang, "A Fuzzy Controller with Supervised Learning Assisted Reinforcement Learning Algorithm for Obstacle Avoidance," *IEEE Transactions on Systems, Man*, *and Cybernetics*, *Part B*: *Cybernetics*, Vol. 33, No. 1, 2003, pp. 17-27. doi:10.1109/TSMCB.2003.808179

[22] M. J. Er and C. Deng, "Obstacle Avoidance of a Mobile Robot Using Hybrid Learning Approach," *IEEE Transactions on Industrial Electronics*, Vol. 52, No. 3, 2005, pp. 898-905. doi:10.1109/TIE.2005.847576

[23] M. T. Rosenstein and A. G. Barto, "Learning and Approximate Dynamic Programming: Scaling Up to the Real World," IEEE Press and John Wiley & Sons, Inc., New York, 2004.

[24] J. A. Clouse and P. E. Utgoff, "A Teaching Method for Reinforcement Learning," *Proceedings of the Nineth International Conference on Machine Learning*, San Francisco, 1992, pp. 92-101.

[25] H. Benbrahim, and J. A. Franklin, "Biped Dynamic Walking Using Reinforcement Learning," *Robotics and Autonomous Systems*, Vol. 22, No. 3-4, 1997, pp. 283-302. doi:10.1016/S0921-8890(97)00043-2

[26] K. Pathak, J. Franch and S. K. Agrawal, "Velocity and Position Control of a Wheeled Inverted Pendulum by Partial Feedback Linearization," *IEEE Transactions on Robotics*, Vol. 21, No. 3, 2005, pp. 505-513. doi:10.1109/TRO.2004.840905

[27] A Blankespoor and R Roemer, "Experimental Verification of the Dynamic Model for a Quarter Size Self-Balancing Wheelchair," *Proceedings of the* 2004 *American Control Conference*, Boston, 2004, pp. 488-492.

[28] F Grasser, A. D'Arrigo, S. Colombi and A. C. Rufer, "JOE: A Mobile, Inverted Pendulum," *IEEE Transactions on Industrial Electronics*, Vol. 49, No. 1, 2002, pp. 107-114. doi:10.1109/41.982254

[29] D. P. Anderson, "NBot Balancing Robot, a Two Wheel Balancing Robot," 2003. http://www.geology.smu.edu/~dpa-ww/robo/nbot/index.html.

[30] C.-H. Chiu, "The Design and Implementation of a Wheeled Inverted Pendulum Using an Adaptive Output Recurrent Cerebellar Model Articulation Controller," *IEEE Transactions on Industrial Electronics*, Vol, 57, No. 5, 2010, pp. 1814-1822. doi:10.1109/TIE.2009.2032203

[31] S. Jung and S. S. Kim, "Control Experiment of a Wheel-Driven Mobile Inverted Pendulum Using Neural Network," *IEEE Transactions on Control Systems Technology*, Vol. 16, No. 2, 2008, pp. 297-303. doi:10.1109/TCST.2007.903396

[32] L. Q. Han and X. Y. Tu, "Study of Artificial Brain Based on Multi-Centrum Self-Coordination Mechanism," Science Press, Beijing, 2009.

[33] C. Chen and R. F. Thompson, "Temporal Specificity of Long-Term Depression in Parallel Fiber - Purkinje Synapses in Rat Cerebellar Slice," *Learning and Memmory*, Vol. 2, No. 3-4, 1995, pp. 185-198. doi:10.1101/lm.2.3-4.185

[34] B. F. Skinner, "The Behavior of Organisms," Appleton-Century-Crofts, New York, 1938.

[35] I. P. Pavlov, "Conditioned Reflexes," Oxford University Press, Oxford, 1927.

[36] B. Brembs, W. Plendl, "Double Dissociation of PKC and AC Manipulations on Operant and in Drosophila," *Current Biology*, Vol. 18, No. 15, 2008, pp. 1168-1171. doi:10.1016/j.cub.2008.07.041

[37] C. W. Yao and G. C. Chen, "A Emotion Development Agent Model Based on OCC Model and Operant Conditioning," 2001 *International Conferences on Info-Tech and Info-Net Proceedings*, Beijing, 2001, pp. 246-250.

[38] K.Itoh, H. Miwa, M. Matsumoto, M. Zecca, M. Takanobu, H. Roccella, S. Carrozza, M. C. Dario and P. Takanishi, "Behavior Model of Humanoid Robots Based on Operant Conditioning," *Proceedings of* 2005 5*th IEEE-RAS International Conference on Humanoid Robots IEEE/RAS International Conference on Humanoid Robots*, Tsukuba, 2005, pp. 220-225.

[39] J. S. Leng, L. Jain and C. Fyfe, "Convergence Analysis on Temporal Difference Learning. International Journal

of Innovative Computing," *Information and Control*, Vol. 5, No. 4, 2009, pp. 913-922.

[40] X. G. Ruan, "Neural Computational Science: Simulation Brain Function at the Cellular Level," National Defense Industry Press, Beijing, 2006.

[41] M. Murata and S. Ozawa, "A Reinforcement Learning Model Using Deterministic State-Action Sequences," *International Journal of Innovative Computing, Information and Control*, Vol. 6, No. 2, 2010, pp. 577-590.

[42] J. Randlov, A. G. Barto and M. T. Rosenstein, "Combining Reinforcement Learning with a Local Control Algo-rithm," *Proceedings of the Seventeenth International Conference on Machine Learning*, Vol. 1, No. 4, 2000, pp. 775-782.

[43] C. T. Chiang and C. S. Lin, "CMAC with General Basis Functions," *Neural Network*, Vol. 9, No. 7, 1996, pp. 1199-1211. doi:10.1016/0893-6080(96)00132-3

[44] A. V. Lazo and P. Rathie, "On the Entropy of Continuous Probability Distributions," *IEEE Transactions on Information Theory*, Vol. 24, No. 1, 1978, pp. 120-122. doi:10.1109/TIT.1978.1055832

Appendix

The detail dynamic model of self-balancing two-wheeled robot is as follows.

$$M(\tilde{q})\ddot{\tilde{q}} + N(\tilde{q}, \dot{\tilde{q}}) = ET \tag{17}$$

where,

$$M(\tilde{q}) = \begin{bmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{bmatrix};$$

$$N(\tilde{q}, \dot{\tilde{q}}) = \begin{bmatrix} n_1 \\ n_2 \\ n_3 \end{bmatrix}; E = \begin{bmatrix} -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$m_{11} = \left(ML^2 + J_y\right)$$

$$m_{21} = m_{31} = m_{12} = m_{13} = \frac{1}{2}MLR\cos\theta_b$$

$$m_{22} = \left(m_l R^2 + J_w\right) + \frac{1}{4}MR^2 + \frac{R^2}{d^2}\left(ML^2\sin^2\theta_b + J_z\right)$$

$$m_{23} = m_{32} = \frac{1}{4}MR^2 - \frac{R^2}{d^2}\left(ML^2\sin^2\theta_b + J_z\right);$$

$$m_{33} = \left(m_r R^2 + J_w\right) + \frac{1}{4}MR^2 + \frac{R^2}{d^2}\left(ML^2\sin^2\theta_b + J_z\right)$$

$$n_1 = -MgL\sin\theta_b - M\dot{\theta}^2 L^2 \sin\theta_b \cos\theta_b + 2\mu_w\dot{\theta}_b - \mu_w\left(\dot{\varphi}_l + \dot{\varphi}_r\right)$$

$$n_2 = ML^2\left(\frac{R}{d}\right)^2\left(\dot{\varphi}_l - \dot{\varphi}_r\right)\dot{\theta}_b\sin 2\theta_b - \frac{1}{2}MLR\dot{\theta}_b^2\sin\theta_b + \left(\mu_l + \mu_w\right)\dot{\varphi}_l - \mu_w\dot{\theta}_b$$

$$n_3 = -ML^2\left(\frac{R}{d}\right)^2\left(\dot{\varphi}_l - \dot{\varphi}_r\right)\dot{\theta}_b\sin 2\theta_b - \frac{1}{2}MLR\dot{\theta}_b^2\sin\theta_b + \left(\mu_r + \mu_w\right)\dot{\varphi}_r - \mu_w\dot{\theta}_b$$