Brain-inspired Motion Learning in Recurrent Neural Network with Emotion Modulation

Xiao Huang, Wei Wu, Hong Qiao, *Senior Member, IEEE*, and Yidao Ji

***Abstract*—Based on basic emotion modulation theory and the neural mechanisms of generating complex motor patterns, we introduce a novel emotion-modulated learning rule to train a recurrent neural network, which enables a complex musculoskeletal arm and a robotic arm to perform goal-directed tasks with high accuracy and learning efficiency. Specifically, inspired by the fact that emotions can modulate the process of learning and decision making through neuromodulatory system, we present a model of emotion generation and modulation to adjust the parameters of learning adaptively, including the reward prediction error, the speed of learning and the randomness in action selection. Additionally, we use Oja learning rule to adjust the recurrent weights in delayed-reinforcement tasks, which outperforms the Hebbian update rule in terms of stability and accuracy. In the experimental section, we use a musculoskeletal model of the human upper limb and a robotic arm to perform goal-directed tasks through trial-and-reward learning respectively. The results show that emotion-based methods are able to control the arm with higher accuracy and a faster learning rate. Meanwhile, emotional Oja agent is superior to emotional Hebbian one in term of performance.**

***Index Terms*—Brain-inspired Model, Emotion, Recurrent Neural Network, Motion Learning.**

I. Introduction

In recent years, there is an expansive research on biologically-inspired models for solving complex cognition, decision making and motor control problems in the field of robotics. Many robots have been designed to mimic human behavior, which requires the integration of various technologies such as robotics, neuroscience and artificial intelligence. Learning to perform diverse tasks of computation and pattern generation is still a huge challenge for a humanoid robot. For example, the muscular system is one of the major actuator modules of animals. Many recent studies have used the biomimetic technology to develop some bio-inspired muscle hardware for mimicking animal's movement. However, due to the strong redundancy and coupling, generating complex patterns of muscle activities is difficult for a robot system actuated by artificial muscles.

Many methods aim to solve this kind of control problem in the areas of robotics [1]–[3]. Some compute a set of muscle excitations to drive a dynamic musculoskeletal model towards a desired kinematic trajectory through the *Computed Muscle Control*, which use a combination of PD control and static optimization [4]. However, there are some disadvantages to this approach. On the one hand, the desired kinematic trajectory must be planned before computation. On the other hand, the computational expense of this method is generally high. Another way is to use reinforcement learning. Recent works tries to develop a controller to enable a human musculoskeletal model to run in a complex obstacle course as quickly as possible. Deep Reinforcement Learning is brought to solve this problems in medicine. Despite of success in the complex obstacle course, the methods generally need to construct a very high-dimensional input data, and the computation of them is usually expensive. Meanwhile, These methods are not enough biologically plausible. Human brain is a highly dynamic system with lots of recurrent connections, and each neuron obeys a specific non-linear dynamic function. Here, we aim to apply a biologically-plausible recurrent neural network to perform the complex decision making tasks.

Recurrent neural networks (RNN) have been considered as biologically plausible models for simulating populations of neurons because of their complicated dynamics and highly recurrent connections. Populations of neurons in cerebral cortex play an important role in cognitive computing and motor learning. A number of researches in computational neuroscience have found that populations of neurons can optimize some specific objective functions during learning [5]–[7]. Thus, lots of recurrent neural networks is utilized to model the process of biologically relevant computations, such as generating coherent patterns of activity [8], implementing flexible decision making, motor control, associations and memory maintenances [9], [10].

For training these networks, there are two classes of methods: supervised learning [8], [11], [12] and reinforcement learning [9], [13]. A range of supervised RNN training methods have been proposed from the classical backpropagation to reservoir methods: *Backpropagation Through Time* [12], *Atiya-Parlos recurrent learning* [14], *BackPropagation-DeCorrelation* [15] and *Echo State Networks* [16]. However, these supervised training methods generally require a derivable objective function or a continuous supervisory signal, and weights of the neural networks are usually optimized by gradient decent method. In contrast, reinforcement learning

H. Qiao, W. Wu, X. Huang are with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China, and the University of Chinese Academy of Sciences, e-mail: (hong.qiao@ia.ac.cn).

H. Qiao is with the CAS Center for Excellence in Brain Science and Intelligence Technology, and the Cloud Computing Center, Chinese Academy of Sciences.

Y. Ji is with School of Automation and Electrical Engineering, University of Science and Technology, Beijing, China.

This work is supported by the development of science and technology of Guangdong province special fund project (Grant 2016B090910001), the Strategic Priority Research Program of the CAS (Grant XDB02080003), the Beijing Municipal Science and Technology (Grants D16110400140000 and D161100001416001), the National Natural Science Foundation of China (Grant 61210009, 61627808, 91648205, 51705515, U1713201 and 61702516).

is more biologically plausible. This method is inspired by the modulation effects of dopamine on synaptic plasticity in neuroscience, where the changes of synaptic weights depend on not only the activities of pre-synaptic and post-synaptic neurons, but also the reward signals [17]. Recent deep reinforcement algorithms generally add exploratory noise on the space of action directly, which however causes a very low convergence speed in the muscle control problem. Meanwhile, it is difficult to design and explain the structure of neural networks, such as the size of them. A large-size network is usually unstable or hard to train, and a small network is insufficient to deal with the complex learning processes. *Evolution strategies* is an alternative to reinforcement learning, where the noise is added to the weights of neural networks for improving the performance of exploration. However, this evolution algorithm is unsuitable for performing some time-sensitive tasks, and commonly suffers from the expensive cost of computation.

We use a fully connected recurrent neural network to simulate populations of neurons in motor cortex. Each neuron is a leaky integrator, which describes the dynamics of neural membrane potential. Node-perturbation method [18], [19] is applied to train this neural network, where the exploratory noise is added to the membrane potential of neurons directly. Different from perturbation in action space or weight space, node-perturbation method regards the exploratory noise as disturbing currents in neural activities. We propose a novel form of reward-modulated Oja learning to adjust the weights of the network. Our learning rule can reduce the dimensionality of weight space being searched, which allows the network to reach an optimal state sufficiently quickly. Compared with traditional Hebbian learning, this method is insensitive to the size of network. Meanwhile, emotion modulation is integrated into the process of motor learning. Emotions are one of the most powerful determinant factors of our perception, cognition, decision-making and other behaviors. Some works consider that emotions can influence decision making through reward modification, state modification, meta-learning and action selection in reinforcement learning [20]. We follow the hypothesis that neuromodulatory systems are the media for dynamically adjusting metaparameters of reinforcement learning in the brain [21]. In this paper, the learning parameters are adaptively adjusted by emotional valence that is computed by the entropy of the rewards over time. A series of simulated and real robotic experiments have demonstrated that our proposed method is able to improve the performance of robotic learning effectively.

The rest of this paper is organized as follows: section II firstly presents the related biological and psychological background of emotion, including the circuits in emotion generation, regulation and modulation on decision making. And the biological findings of reward-based learning and motor learning are also argued further. In section III, several computational architectures of emotion generation and emotional decision-making processes are proposed to facilitate theoretical analysis of emotion-based motor learning. After that, a novel emotional reward-modulated learning is presented to train the recurrent neural network for generating complex patterns of movements. In section IV, our algorithm is implemented on a musculoskeletal model of the human upper limb and a real robotic arm respectively. The experimental results are discussed. Finally, conclusions are given in section V.

II. Background

A. Generation and Representation of Emotion

In this paper, emotion is modelled and integrated into a biologically-inspired model to mimic the modulation on decision making. Before that, the related evidences or findings of emotion in neuroscience and psychological studies have been reviewed and discussed here.

1) Generation and regulation of emotion: Emotion is one of the characteristic features of human, which consists of curiosity, happy, fear, stress, anxiety and etc. Research on emotion has been a hot topic in the field of neuroscience for decades. Generally, emotions arise from an interaction between person and situation, which has a valenced meaning to an individual [22]. Some researches in cognitive psychology consider that emotions are generated and regulated via the interaction of bottom-up and top-down processes [23]. Bottom-up processes can produce quick affective analyses of low-level perceptual stimuli. While top-down processes can regulate what emotions we have according to high-level top-down cognitive appraisal processes. In the neural systems, the orbitofrontal cortex (OFC), hippocampus, striatum, insular cortex and amygdala are important in emotion processing [24]–[27]. For example, amygdala is crucial for the fear learning [28]. The feedforward projections from amygdala are terminated in the middle layers of OFC, where the significance of the stimuli may be conveyed [29].

Emotional state is often adjusted to suit our needs over time, where regulation of emotion plays an important role to modify the generative processes of emotion. According to the biological mechanisms of emotion processing [24]–[27], the refined emotions are combination of fast low-level automatic emotional response and high-level cognitive evaluation. When perceptual stimuli come up to emotion-related regions, quick low-level processes generate primary emotions automatically [22], [23]. As the situation changes over time, the high-level cognitive regions such as prefrontal cortex could send emotion-regulatory signal to regulate the activities of emotion-related areas so that emotional state is adjusted further.

2) Computational models of emotion: Driven by basic researches on emotions and their potential in other fields such as computational science, an expansive research on computational models of emotion processing has been seen in recent years. Several significant models have emerged in the history of basic emotion theory. As a predominant force among psychological perspectives on emotion, cognitive appraisal theory is one of the most fruitful source for design of AI systems. Because in this theory, emotion is argued to arise from individual judgment concerning the relationship between person and environment [30], which are easily associated with the interaction between the agent and environment. Here, appraisal is seen as a main cause of emotion, or at least of the behavioral and cognitive changes associated with emotion [31].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2018.2843563, IEEE Transactions on Cognitive and Developmental Systems

Another popular model is dimensional theory, in which emotional states are composed of points in a continuous dimensional space rather than discrete entities [31]. For example, PAD model is a typical representation of dimensional theory, where every emotional state is built on three basic dimensions corresponding to pleasure, arousal and dominance. However, many dimensional models are based on psychological perspectives instead of biological basis. Recent years have seen a new explanatory model for emotion and monoaminergic neurotransmitters [32], where emotions are generated from the joint modulation of different levels of monoamines, including serotonin, dopamine, noradrenaline. In that work, authors discovered that the eight basic emotions could fit rather well into the eight corners of the cube space spanning from three types of monoamine (Fig.1).

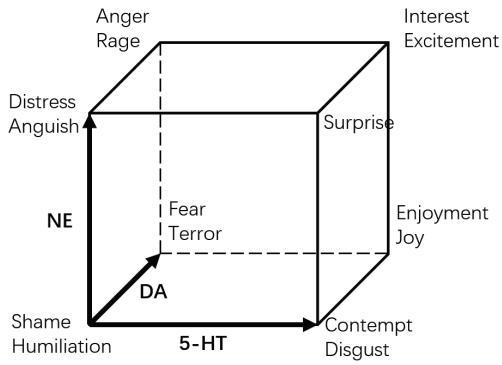

Fig. 1. A three-dimensional model for emotions and monoamine neurotransmitters [32].

B. Modulation of emotion on decision making

A specific affective state could motivate a specific choice in some cases of biological reflexes. For example, the amygdala is crucial for the evaluation of threats and aversive stimuli [33], [34]. The lateral nucleus in amygdala sends signals to central nuclei, which projects to the hypothalamus and brain stem nuclei to mediate the threat response. Moreover, the lateral nucleus also projects to basal nuclei and then send information to striatum, which integrates motivation with action values, and is important for avoidance actions [35]. In case of threat, this circuit plays an important role for avoidance actions in decision-making process.

The emotional state could modulate decision-making process through influencing the computation of subjective value of choices. For example, automatic emotional assessment may be performed by the anterior cingulate cortex (ACC) in the action-monitoring processes during conflict [36]. In the case of conflict, ACC could drive the subthalamic nucleus (STN) to inhibit fast action selection in basal ganglia and allow for consideration of benefits [37]; while in the absence of conflict, the emotional reward-related learning process will take over [38], [39]. Briefly, implementing actions from decision leads to positive/negative emotional error assessment, which will further trigger the mid-brain dopaminergic signaling (mainly from ventral tegmental area) to increase/decrease the probability of future choices [40].

For the computational implementation, several studies have integrated the computational models of emotion into reinforcement learning to explain observed behavior or underlying neurobiological processes. For instance, inspired by the fact that different monoamines can regulate the learning parameters in our brain, the work [21], [41] has presented a hypothesis that neuromodulatory systems are the media for dynamically adjusting metaparameters of learning in the brain. More specifically, it is stated that dopamine and serotonin mainly influences the reward prediction error, noradrenaline affects the randomness in action selection, and acetylcholine controls the speed of memory update. Moreover, the neural bases of emotion regulation are recently drawn into value-based decision making and reinforcement learning framework, where an emotion is described as a perception-valuation-action sequence [42]. Within this framework, developing computational model of modulation of emotion on decision making and reinforcement learning may now become possible.

C. Reward-based learning

Rewards are widely involved in reinforcement learning, motivated behavior, economic choices and generation of emotion. All basal ganglia nuclei or other brain areas associated with basal ganglia have been investigated as a core role in reward-based learning, including reward prediction errors and action value coding in reward-related activities [43]. For example, a range of neuroscience experiments demonstrate that three major brain structures associated with basal ganglia, namely midbrain dopamine neurons, pedunculopontine nucleus and striatum, perform a variety of reward functions. Specifically, midbrain dopamine neurons are able to code a reward prediction error that measures the difference between received and predicted reward, which generally provide teaching signals in temporal difference reinforcement learning model [44]. The dopamine reinforcing signal, as a third factor, can modulate the synaptic transmission via Hebbian plasticity [40], [45], [46]. Positive reinforcing signal would enhance behavior-related neural activity, motivating a agent to perform behavior with increasing reward. Whereas negative reinforcing signal would restrain the agent from doing something that lead to diminished reward. The pedunculopontine nucleus, projecting to dopamine neurons, is considered to code the possible contribution of input sensor, movement or reward stimuli without reward-predicting error information [43]. While neurons in striatum are mainly responsible for integrating motor and reward processing, including action value coding and changing reward-related activity during learning [43]. Unsurprisingly, these neurons coding action values provide important cues in neuronal decision making process, where the action related to highest value would be selected.

III. Method

A. System Architecture

As mentioned above, emotion can influence decision making through modulating a range of high-level cognitive functions. When people face choices, they generally evaluate the possible consequences and choose the option with highest

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2018.2843563, IEEE Transactions on Cognitive and Developmental Systems

value. Meanwhile, in some cases, a specific affective state will directly lead to a specific choice. Here, a feedforward model of emotion affecting decision making is proposed to illustrate two aspects of emotion-based decision making. First, emotional responses can directly trigger emergent action selection and execution (fight-or-flight in face of danger). Second, emotion can modulate higher level cognitive functions (such as attention, working memory and affective memory) to make proper choices in accordance with present situation [47], [48]. A schematic diagram of feedforward emotional decision making is shown in Fig.2.

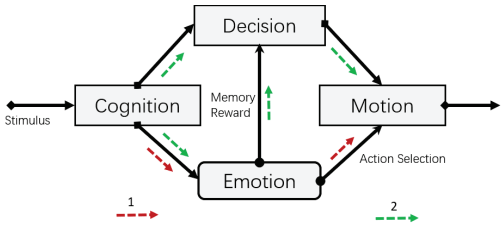

Fig. 2. The feedforward model of emotion-induced decision making. The red arrows represent the quick pathway, and the green arrows denote the slow pathway of emotion modulation.

Moreover, emotions also contribute to evaluate the subjective value of decision-making choices according to the person-environment interaction [42], [48]. As intrinsic reward signals, emotions are generated by evaluating the difference between actual results and individual expectations. Based on the mechanisms mentioned above, a feedback model of emotion-based decision making is built to extend the computational architecture. A schematic diagram of the feedback model of emotion-based decision making is shown in Fig.3.

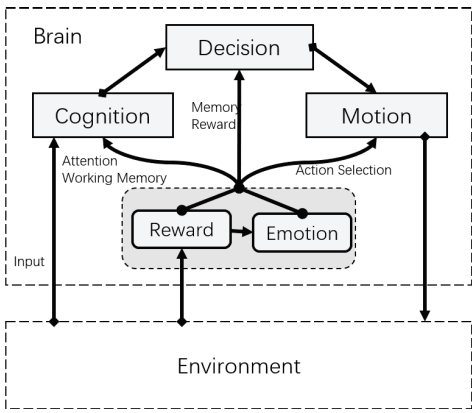

Fig. 3. The feedback model of emotion-modulated learning. Emotions, as intrinsic reward signals, are generated by evaluating the extrinsic rewards. They can modulate high-level cognition, decision-making process and the execution of the actions.

We presented a hypothesis that emotions are able to dynamically modulate the parameters of reward-based learning or other decision making process through neuromodulatory systems. Neuromodulators are the media for regulating the specific global variables such as the speed of learning, the noise for exploration. According to the explanatory model for emotion and monoaminergic neurotransmitters [32], emotions are generated from the joint modulation of different levels of neuromudulators such as serotonin, dopamine, noradrenaline. While such different neuromudulators would mediate learning parameters further, thus we propose the following hypothesis to explain the role of emotion on reward learning or other decision making process. (1) Emotion influences the reward prediction error by adjusting the reward prediction baseline, inspired by dopaminergic modulation in the cerebral cortex; (2) Emotion modulates the speed of learning, which derives from that acetylcholine modulates the synaptic plasticity in the hippocampus and the cerebral cortex; (3) Emotion changes the randomness in action selection by controlling the size of noise for exploration, as noradrenaline balances exploration and focused execution in the brain.

B. Neural Network

In the control of goal-directed and reward-oriented behavior, selecting a proper computing model to generate different motor patterns is a key problem. To implement a wide range of spatiotemporal tasks usually requires encoding time into the dynamic changes of activity of neurons. As mentioned above, population of neurons is an effective neural structure to encode information for neural computations, where the activation of neurons at any given time corresponds to specific points in a high-dimensional space. As a result, this neural network shows complex characteristic of dynamics and highly reminiscent of neural activity so that dealing with a range of spatiotemporal processes becomes feasible. In this paper, a biologically plausible recurrent neural network operating in the near-chaotic regime is applied to model the population of neurons.

Here, we use a fully connected recurrent neural network with N leaky integrator neurons. In each time step, a leaky integration of neural activation is performed from previous time steps. Specifically, the membrane potential x_i of each neuron i is governed by following dynamic equation.

$$\tau \dot{x}_i = -x_i + \sum_{j=1}^{N} W_{ij}^{rec} r_j + \sum_{k=1}^{M} W_{ik}^{in} u_k + \xi_i \tag{1}$$

$$r_i = \tanh(x_i) \tag{2}$$

$$z_l = \sum_{i=1}^{N} W_{li}^{out} r_i \tag{3}$$

where τ is a time constant of postsynaptic neuron. x_i refers to the membrane potential of neuron i, r_i is its firing rate. W_{ij}^{rec} is the synaptic weight from presynaptic node j to postsynaptic node i in recurrent network, similarly, W_{ik}^{in} is the weight of connection from input stimulus u_k to recurrent node i and W_{li}^{out} is the output weight from node i to readout unit z_l. In addition, ξ_i denotes the perturbation of membrane potential of node, which is supposed to conform to normal distribution with zero mean and σ^2 variance. In node-perturbation method [19], these noises could lead to some exploratory variation of action selection that results in different evaluation of each

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2018.2843563, IEEE Transactions on Cognitive and Developmental Systems

trial, and then an optimal strategy would be found through estimating the policy gradient in the perspective of evolutionary computation. If noises are set to zero, the network won't have the learning ability.

In practice, the continuous-time RNN is usually performed in discrete form. Here, we rewrite the above RNN with a conciser expression, where $\boldsymbol{x}$ denotes the input stimuli to the recurrent nodes, $\boldsymbol{r}$ and $\boldsymbol{h}$ are hidden variables, and $\boldsymbol{o}$ refers to the output of neural network.

$$\boldsymbol{r}_t = (1-\alpha)\boldsymbol{r}_{t-1} + \alpha(\boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{W}\boldsymbol{h}_{t-1}) + \boldsymbol{\xi} \tag{4}$$

$$\boldsymbol{h}_t = \tanh(\boldsymbol{r}_t) \tag{5}$$

$$\boldsymbol{o}_t = \boldsymbol{V}\boldsymbol{h}_t \tag{6}$$

where $\alpha = \triangle t/\tau$.

C. Learning Rule

In most actual cognitive and decision-making tasks, a derivable cost function or a continuous supervisory signal is generally absent. Especially in some reinforcement learning tasks, only a sparse and delayed reward is delivered into the learning system after each trial. As for this case, the most relevant learning method is the associative Hebbian learning rule. In 1949, Hebb [49] proposed that if the pre-synaptic and post-synaptic neurons are activated together, then the synaptic strength could be changed. It is often generalized as

$$\triangle w_{ij}(t) = \eta(t) f(y_j(t), x_i(t)) \tag{7}$$

$$\triangle w_{ij}(t+\triangle t) = w_{ij}(t) + \triangle w_{ij}(t) \tag{8}$$

where $f(\cdot,\cdot)$ is a function of pre-synaptic signal $x_i(t)$ and post-synaptic signal $y_j(t)$. $\eta(t)$ is the learning rate at time step t, w_{ij} is the weight of connection from neuron j to neuron i, $x_i(t)$ is the input for neuron i, and $y_j(t)$ is the postsynaptic response.

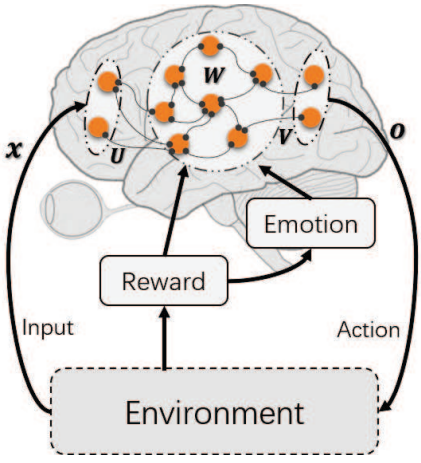

Fig. 4. The structure of emotion-modulated learning.

Based on the associative Hebbian learning rule, a class of associative reinforcement learning, called REINFORCE algorithms, is proposed to make the weight adjustment along the gradient of expected reinforcement in both immediate-reinforcement tasks and delayed-reinforcement tasks [50]. Largely similar to the REINFORCE algorithm, a form of reward-modulated Hebbian learning recently was applied to train a recurrent neural network successfully [13], [51]. This rule builds on the node-perturbation method [19] that estimates the gradient of an objective function by measuring the fluctuations in the objective function in response to dynamic perturbations of neuron activities. Here, node-perturbation learning is proposed to train the RNN mentioned above for delayed-reinforcement tasks.

Suppose that there exits a set of optimal parameters $\boldsymbol{\theta}^*$ of the network such that $\boldsymbol{d} = g(\boldsymbol{x} \mid \boldsymbol{\theta}^*)$, which is defined as a teacher network. The purpose of learning is to train the network to produce desired corresponding outputs $\boldsymbol{o}^0 = \boldsymbol{d}$, where $\boldsymbol{o}^0 = g(\boldsymbol{x} \mid \boldsymbol{\theta})$ refers to outputs of noiseless network. The squared error function is employed to evaluate the performance.

$$E^0 = \frac{1}{2} \parallel f(\boldsymbol{o}^0) - f(\boldsymbol{d}) \parallel^2 \tag{9}$$

where $f(\cdot)$ is the function of system affected by the neural network.

Node-perturbation approach is to add the noise $\boldsymbol{\xi}$ to all neural nodes, and each element of $\boldsymbol{\xi}$ is subject to a Gaussian distribution with zero mean and variance σ^2. And then, the new squared error function of noisy network is

$$E^\xi = \frac{1}{2} \parallel f(\boldsymbol{o}^\xi) - f(\boldsymbol{d}) \parallel^2 \tag{10}$$

where $\boldsymbol{o}^\xi = g(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\xi})$.

The recurrent neural network is trained from trial to trial, where each trial consists of T time steps. According to episodic REINFORCE algorithm [50], if a reinforcement value $R^\xi = -E^\xi$ is delivered to the network at the end of each episode, the weights of connection would be updated as

$$\triangle \boldsymbol{\theta} = \eta \sum_{t=1}^{T} \frac{\partial \mathbb{E}[R^\xi \mid \boldsymbol{\theta}]}{\partial \boldsymbol{\theta}} \tag{11}$$

where η refers to the rate of learning. The derivative with respect to weights $\boldsymbol{U}$ and $\boldsymbol{W}$ is the characteristic eligibility, which can be given as

$$\begin{aligned} \frac{\partial \mathbb{E}[R^\xi \mid \boldsymbol{\theta}]}{\partial \boldsymbol{U}} &= \frac{\partial \mathbb{E}[R^\xi \mid \boldsymbol{\theta}]}{\partial \boldsymbol{o}_t^\xi} \frac{\partial \boldsymbol{o}_t^\xi}{\partial \boldsymbol{h}_t^\xi} \frac{\partial \boldsymbol{h}_t^\xi}{\partial \boldsymbol{r}_t^\xi} \frac{\partial \boldsymbol{r}_t}{\partial \boldsymbol{U}} \\ &= \frac{\partial \mathbb{E}[R^\xi \mid \boldsymbol{\theta}]}{\boldsymbol{r}_t^\xi} \frac{\partial \boldsymbol{r}_t}{\partial \boldsymbol{U}} \\ &= (R^\xi - R^0) f(\boldsymbol{r}_t^\xi, \boldsymbol{r}_t^0) \boldsymbol{x}_t^{\mathrm{T}} \end{aligned} \tag{12}$$

$$\begin{aligned} \frac{\partial \mathbb{E}[R^\xi \mid \boldsymbol{\theta}]}{\partial \boldsymbol{W}} &= \frac{\partial \mathbb{E}[R^\xi \mid \boldsymbol{\theta}]}{\partial \boldsymbol{o}_t^\xi} \frac{\partial \boldsymbol{o}_t^\xi}{\partial \boldsymbol{h}_t^\xi} \frac{\partial \boldsymbol{h}_t^\xi}{\partial \boldsymbol{r}_t^\xi} \frac{\partial \boldsymbol{r}_t}{\partial \boldsymbol{W}} \\ &= \frac{\partial \mathbb{E}[R^\xi \mid \boldsymbol{\theta}]}{\boldsymbol{r}_t^\xi} \frac{\partial \boldsymbol{r}_t}{\partial \boldsymbol{W}} \\ &= (R^\xi - R^0) f(\boldsymbol{r}_t^\xi, \boldsymbol{r}_t^0) \boldsymbol{h}_{t-1}^{\mathrm{T}} \end{aligned} \tag{13}$$

where $f(\boldsymbol{r}_t^\xi, \boldsymbol{r}_t^0)$ is applied to estimate the deviations between the noisy and noiseless traces of hidden variable $\boldsymbol{r}$. To learn from sparse, delayed rewards, [9] proposed that this term is a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2018.2843563, IEEE Transactions on Cognitive and Developmental Systems

supralinear function for amplifying the larger deviations and suppressing small ones. The choice of this function is not crucial as long as it is supralinear. Here a cubic function is used.

$$f(\boldsymbol{r}_t^{\xi}, \boldsymbol{r}_t^0) = (\boldsymbol{r}_t^{\xi} - \boldsymbol{r}_t^0)^3 \tag{14}$$

Since $\boldsymbol{r}^0$ is highly related to previous perturbation rather than constant in the case of strongly recurrent network, the deviations between the noisy and noiseless activation couldn't be computed directly. To address this problem, an approach is to predict the mean of activation and regard it as the baseline from noiseless network. Generally, the mean can be estimated by the moving average of the actual values $\boldsymbol{r}_t^{\xi}$.

$$\boldsymbol{r}_t^0 = \alpha^r \boldsymbol{r}_{t-1}^0 + (1-\alpha^r)\boldsymbol{r}_t^{\xi} \tag{15}$$

Then, the weight matrix of input and recurrent connection can be changed incrementally based on following rule.

$$\triangle \boldsymbol{U} = \eta(R^{\xi} - R^0)\sum_{t=1}^{T}(\boldsymbol{r}_t^{\xi} - \boldsymbol{r}_t^0)^3 \boldsymbol{x}_t^{\mathrm{T}} \tag{16}$$

$$\triangle \boldsymbol{W} = \eta(R^{\xi} - R^0)\sum_{t=1}^{T}(\boldsymbol{r}_t^{\xi} - \boldsymbol{r}_t^0)^3 \boldsymbol{h}_{t-1}^{\mathrm{T}} \tag{17}$$

However, the Hebbian update rule mentioned above has a severe problem: sometimes, the synaptic weights may become very large due to lack of something to balance the growth of these weights. Meanwhile, if the network is large, searching a set of optimal weights becomes particularly difficult because the dimensionality of the parameter space is very high. Oja learning rule use a forgetting term to restrain the unlimited growth of weights [52], [53]. The forgetting term is proportional to the product of the synaptic weights and the square of post-synaptic signal. There is a significant property for Oja learning that the final synaptic weights are proportional to the principal component of the input-unit correlation matrix. Thus, this learning rule actually can reduce the dimensionality of weight space to the dimension of the eigenvector of the input-unit correlation matrix. If the dimension of input unit is low, the weight space being searched will become smaller so that the weights are easier to reach the optimal value. Based on this, we propose that the recurrent weights are modified by Oja update rule for a faster learning speed. Formally, the update function is defined as follows.

$$\triangle \boldsymbol{W} = \eta(R^{\xi} - R^0)(\sum_{t=1}^{T} \boldsymbol{e}_t^w - (R^{\xi} - R^0)\boldsymbol{W}) \tag{18}$$

where $\boldsymbol{e}_t^w = (\boldsymbol{r}_t^{\xi} - \boldsymbol{r}_t^0)^3 \boldsymbol{h}_{t-1}^{\mathrm{T}}$ is the characteristic eligibility of recurrent weight at time step t, which represents a potential modification of synaptic weight.

The baseline R^0 is the reward for the noiseless network, which also refers to an adaptive estimate of upcoming reward based on past experience, so that $R^{\xi} - R^0$ represents the reward prediction error signal. A general simple approach to compute the value of reward prediction online is to maintain a running average of actual rewards [54]. So at the n^{th} episode, the reward prediction is given by

$$R_n^0 = \alpha^R R_{n-1}^0 + (1-\alpha^R)R_n^{\xi} \tag{19}$$

where $0 < \alpha_r \leq 1$.

Similarly, as for the weight of output connection, the characteristic eligibility of output weight $\boldsymbol{V}$ is given as follows.

$$\begin{aligned} \frac{\partial \mathbb{E}[R^{\xi} \mid \boldsymbol{\theta}]}{\partial \boldsymbol{V}} &= \frac{\partial \mathbb{E}[R^{\xi} \mid \boldsymbol{\theta}]}{\partial \boldsymbol{o}_t^{\xi}} \frac{\partial \boldsymbol{o}_t^{\xi}}{\partial \boldsymbol{V}} \\ &= (R^{\xi} - R^0)(\boldsymbol{o}_t^{\xi} - \boldsymbol{o}_t^0)^3 \boldsymbol{h}_t^{\mathrm{T}} \end{aligned} \tag{20}$$

where the mean of outputs can be estimated by the moving average of the actual outputs.

$$\boldsymbol{o}_t^0 = \alpha^o \boldsymbol{o}_{t-1}^0 + (1-\alpha^o)\boldsymbol{o}_t^{\xi} \tag{21}$$

where $0 < \alpha_o \leq 1$. And then, the weight matrix of output connection would be modified incrementally based on following rule.

$$\triangle \boldsymbol{V} = \eta(R^{\xi} - R^0)\sum_{t=1}^{T}(\boldsymbol{o}_t^{\xi} - \boldsymbol{o}_t^0)^3 \boldsymbol{h}_t^{\mathrm{T}} \tag{22}$$

To train this RNN, it requires an objective function including both the error term and some regulation terms. Generally, the error can be measured by computing the error sum of squares of the difference between the target and actual results. The regulation terms are designed to encourage sparse weights or minimize the energy of neural activation. The objective function will be further described in the experimental section.

D. Emotion Modulation

In this section, the emotional valence is related to the entropy of the rewards over time. Entropy in statistical mechanics is a measure of how concentrated a probability distribution. Assume the change of rewards in a specified time period is subject to a probability distribution $p(x)$ on a space X, and its entropy is:

$$H(X) = \int_{x \in X} (-\ln p(x))p(x)dx \tag{23}$$

In our approach, the entropy of rewards represents the average degree of dispersion of agent's rewards. As the reward increases over time, the agent can counter dispersive effects and keep a stable performance, and the higher the valence of agent will become. We thus define the valence of agent as a decreasing function of the estimation of entropy of the rewards over time. Mathematically, this can be implemented with the following rule.

Define the sequence of sparse rewards among trials is $\{R_n\}_{n=1}^{n=N} > 0$, where N is the total numbers of trials. Then the emotional valence at trial n is estimated through the entropy of population of rewards in a sliding window. Assume the samples of rewards obey a gaussian distribution in this window.

$$R_{(n-d):n} \sim \mathcal{N}(\mu_n, \zeta_n^2) \tag{24}$$

where d is the size of the window. μ_n and ζ_n^2 can be obtained by maximum likelihood estimation of these samples. Then, the entropy is a logarithmic measure of the number of states with significant probability of being occupied, which can be computed as follows.

$$H_n = \ln(\zeta_n) + \frac{1}{2}\ln(2\pi e) \tag{25}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2018.2843563, IEEE Transactions on Cognitive and Developmental Systems

After that, the emotional valence derives from the difference between a short-term and a long-term running average of this entropy, where the long-term one implies primary mental expectation. If the short-term one is continuously greater than the long-term baseline for a long time, the value of valence would run high, and vice versa. Formally, we deal with the entropy to generate the emotional valence by following rule.

$$\phi_n^s = (1 - \frac{1}{\tau_\phi})\phi_{n-1}^s + \frac{1}{\tau_\phi}\exp(-H_n) \tag{26}$$

$$\phi_n^l = (1 - \frac{1}{\tau_\phi})\phi_{n-1}^l + \frac{1}{\tau_\phi}\phi_n^s \tag{27}$$

$$\Phi_n = \sum_{i=1}^{n}(\phi_i^s - \phi_i^l) \tag{28}$$

According to the multi-dimensional model for emotion and monoamine neurotransmitters [32], we consider that positive/negative emotions are generated by the joint modulation of different levels of monoamines such as serotonin, dopamine, noradrenaline and acetylcholine. Such different monoamines have been suggested to mediate the global signals that regulate the learning parameters in our brain [21]. According to the viewpoint of [21], a hypothesis has been proposed: dopamine and serotonin mainly influences the reward prediction error, noradrenaline affects the randomness in action selection, and acetylcholine controls the speed of memory update. Here, we suggest the emotional valence is able to influence decision making through affecting reward prediction error, controlling the learning rate and node-perturbation noise. The the entire flow of emotion generation and modulation is presented in Fig.5.

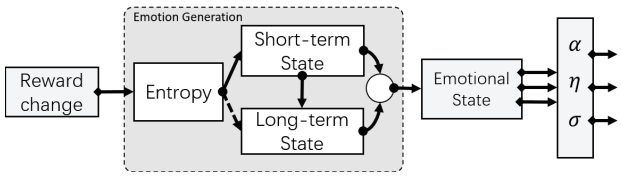

Fig. 5. The schematic diagram of the process of emotion generation and modulation. The dashed line denotes the indirect impacts of entropy on long-term state.

(1) Inspired by dopaminergic modulation in the cerebral cortex, we suggest that emotional valence is able to change the reward prediction error by adjusting the baseline. From the perspective of computation, emotion can adaptively adjust the filter factor so that estimating upcoming reward depends on either more recent rewards or more past experience. Here, the filter factor is simply defined as a decreasing function of valence.

$$\alpha_n^r = \lambda_\alpha \exp(-\frac{\Phi_n}{\tau}) \tag{29}$$

(2) Emotion can modulate the parameter of learning rate η_n. Neurobiological studies [55], [56] have shown that acetylcholine is able to modulate the synaptic plasticity in the hippocampus and the cerebral cortex, and then influences the learning process and memories. Here, emotion modulates

Algorithm 1 Emotion Modulated Decision Making.

Initialize:
Initialize a recurrent neural network.
Training:
for $i = 1$ to N_{trial} **do**
for $t = 1$ to T **do**
Compute the output of network $\boldsymbol{o}_t = g(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\xi})$;
end for
return $\boldsymbol{o} = [\boldsymbol{o}_t]_{t=1}^T$;
Compute the output of system $y = System(\boldsymbol{o})$;
Compute rewards $R = Reward(y, y^t)$ based on an specific objective function;
Compute the value of emotional valence $\Phi = Emotion(R)$;
Adjust the parameters of learning with emotional valence Φ;
Update the weight of network $\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}$.
end for

cognitive and behavioral learning system by controlling the learning rate, which is simply governed by

$$\eta_n = \lambda_\eta \exp(-\frac{\Phi_n}{\tau}) \tag{30}$$

(3) Emotion can control the randomness in action selection through adjusting the variance of the node-perturbation noise. In the viewpoint of neurobiology, noradrenaline has the function of balancing the wide exploration and focused execution. Meanwhile, in related psychological studies [57], [58], on the problems involving risky choices, some specific negative emotions tend to increase risk-taking choices, whereas others act risk-averse. Hence, we propose that the neural system tends to more exploration when valence is low, conversely more exploitation when valence is high, which can be implemented by adjusting the variance of noise.

$$\sigma_n^2 = \lambda_\sigma \exp(-\frac{\Phi_n}{\tau}) \tag{31}$$

In brief, in this section, we respectively introduce the emotional reward-modulated Hebbian and Oja learning rule to fulfil the goal-directed and delayed-reinforcement tasks. As for the entire flow, the recurrent neural network firstly generates an output controlling sequence in response of the corresponding input stimuli in each trial. Then this sequence of control is fed into the actuator, as a result, we can obtain the result of system's output. Based on a certain objective function or reward function from environment, a reward will be returned into the neural system to encourage the neurons to change the synaptic weights. Meanwhile, combining short-term and long-term running average of rewards' entropy is used to generate the emotional valence further. Then the emotional valence can influence the process of learning not only by modulating the reward prediction error, but also changing a series of parameters of learning. By training the neural network from trial to trial, we ultimately make the neural system generate a certain motor pattern sequence as a response to the corresponding input stimuli. The complete algorithm is given in Algorithm 1.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2018.2843563, IEEE Transactions on Cognitive and Developmental Systems

IV. EXPERIMENTS

A. Muscle-skeleton model motion task

To evaluate the effectiveness and efficiency of the proposed algorithm, our learning rule is used to train a recurrent neural network in a classic reaching task. This task, where a monkey moves its hand from a start location to target points, is generally for studying voluntary decision making and motor control. It's also a typical task to study the control of robotics. In the experiment, we build a neuro-muscle-skeleton model to verify the algorithm of emotional motor learning for decision tasks. By continuous trial-and-reward learning, a population of neurons is trained to generate optimal muscle activation signals that control the skeleton to reach specific location in response to the corresponding input signal. Fig.6 is a schematic diagram of this experiment.

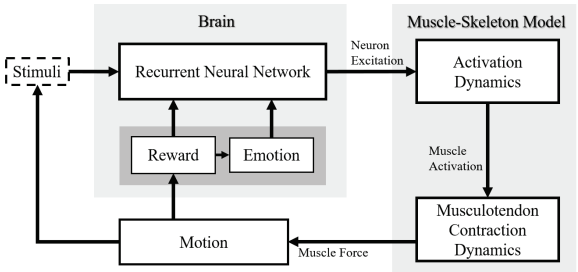

Fig. 6. The structure of neuro-muscle-skeleton model. In each trial, the RNN generates a time series of muscle activation signals according to current stimuli. These signals cause the muscles to change the length of them and pull the skeletons. In the end of each trial, extrinsic reward and intrinsic emotion modulate the learning parameters of the network together.

1) Muscle-Skeleton Model: Here, a musculoskeletal model of human upper limb is chosen based on OpenSim, which is an open source software that allows users to develop models of musculoskeletal structures and create dynamic simulations of movement. We use a modified model described in [9], [59] with the Thelen muscle model [60] to build the upper limb musculoskeletal model as shown in Fig.7. This model with three degrees of freedom at the shoulder and one at the elbow, is actuated by sixteen muscles attached to the chest, shoulder, upper and lower arm bones respectively. The muscle forces can be calculated on the basis of current muscle activation. With the control of all muscle forces, upper-limb bones are able to make a variety of movements [2]. The detailed mathematical description of the muscle activation and force generation could be found in [60].

2) Experimental Design and Results: Our reaching task is to control the fingertip from the starting point $S = (0.01, -0.26, -0.7)$ to one of four target points located at $P_0 = (0.4, 0.2, 0.1)$, $P_1 = (0.4, -0.2, 0.1)$, $P_2 = (0.4, 0.2, -0.3)$, $P_3 = (0.4, -0.2, -0.3)$ respectively. In every moment, the coordinate of target point is fed into the neural network as the input vector. In the experiment, the upper limb executes a movement from starting position to target point within 700 ms. A recurrent neural network with 400 neurons is trained, and the weights of network are updated by the above-mentioned learning rule at the end of each trial. In addition,

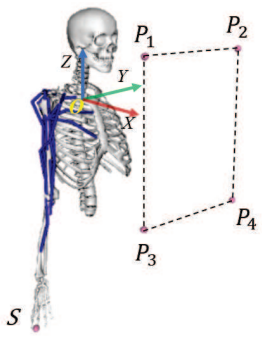

Fig. 7. The musculoskeletal model of human upper limb. The arm is put down without control at the starting position S.

we assume that the fingertip reaches point y with the control of readout activation at each trial, while the target point is y^t. The objective function is composed of an Euclidean distance error and a energy term for muscle activation over the entire trial. As a penalty term, the energy depends on the activation and the max isometric force of muscles.

$$L = -R = \| y - y^t \|_2 + \frac{\lambda}{T} \sum_{j=1}^{T} \sum_{k=1}^{N_{out}} o_{kj} \frac{MIF_k}{\sum_i^{N_{out}} MIF_i} \quad (32)$$

where λ is the parameter of the penalty term, T is total time steps in each trial, N_{out} is the number of output units, o_{kj} represents the activation of muscle k at time step j, and MIF_k is the max isometric force of muscle k.

Properly initializing the parameters of the network can save the learning time. In the experiment, each element of the recurrent weight matrix $\boldsymbol{W}$ is initialized to zero with probability $1-p$, and the remaining entries are set to samples from a Gaussian distribution with zero mean and variance $g/(pN)$ [4]. The elements of input weight $\boldsymbol{U}$ and output weight $\boldsymbol{V}$ conform to a uniform distribution over a small range. Besides, the initial membrane potential x_i of each neuron i also conform to a uniform distribution at the start of each trial. Parameters of the network without emotion modulation are shown in Table.I in detail. For the emotion-modulated learning, emotion-related parameters are given in Table.II.

TABLE I
THE DEFAULT PARAMETERS OF UNEMOTIONAL NEURAL NETWORK

Parameter	Symbol	Default Value
Learning rate	η	0.5
Spectral radius of recurrent weight	g	1.5
Probability of recurrent weight	p	0.5
Number of neurons	N	400
Number of input units	N_{in}	4
Number of output units	N_{out}	16
Unit time constant	τ	30ms
Time step	$\triangle t$	1ms
Variance for recurrent noise	σ	8
Filter factor of reward	α_r	0.33
Parameter of regularization term	λ	0.8

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2018.2843563, IEEE Transactions on Cognitive and Developmental Systems

TABLE II
THE DEFAULT PARAMETERS OF EMOTIONAL NEURAL NETWORK

Parameter	Symbol	Default Value
A short-term and long-term time constant	τ_e	300
Amplitude of learning rate	λ_a	0.6
Amplitude of filter factor	λ_r	0.7
Amplitude of noise variance	λ_s	8

We have conducted three experiments whose learning rules are Oja update without emotion modulation, Hebbian update with emotion modulation and Oja update with emotion modulation respectively. After 5000 trials, their change curves of loss are shown in Fig.8, where the bar of each line is the standard error of the mean. It is clear that performance improves with increasing trials, and error reaches a low value after about 1000 trials. However, compared with emotionless version, the methods with emotion modulation manifest a faster learning process and more stable performance. Specifically, emotion-based methods are able to reach a lower error quickly, which actually benefits from adaptively adjusting the learning parameters through modulation of emotion. In addition, it is obvious that the the emotional Oja agent is superior to the emotional Hebbian one for ensuring higher control precision and searching a more stable policy of motion.

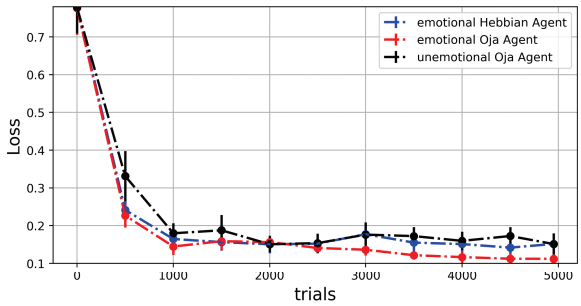

Fig. 8. The loss of the three agents in the musculoskeletal model reaching task, which denotes the sum of positional errors and energy of muscle activation.

After training, 20 tests are conducted to investigate the trajectories of fingertip motion and errors between expected and actual results as shown in Fig.9, where the lines represent the trajectories of fingertip's movements, four red points are targets. Table.III shows the mean and standard deviations of final errors in each pattern of motion about these three agents. Obviously, the emotional agents make the upper limb move smoother and more consistent in each type of test. Among these three agents, emotional Oja agent is able to control the arm with highest accuracy and best stability. It may benefit from the good balancing effect of the forgetting term to the growth of the weight so that very extreme value can not appear. This guarantees the stability of the optimizing algorithm.

We also investigate the generation of emotional valence and the modulation on learning parameters. In the early stage of learning, the rewards change dramatically so that the entropy of population of rewards is high. However, with the better pol-

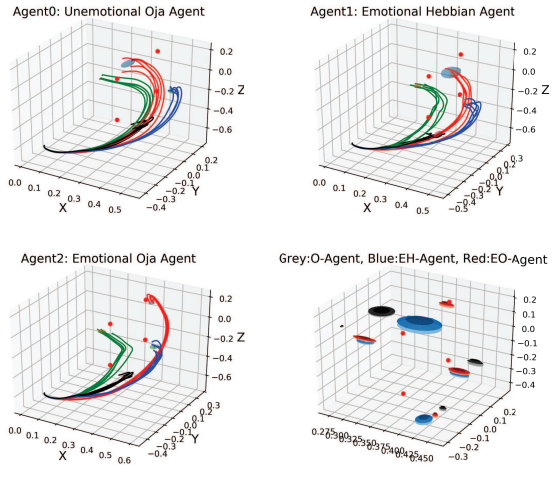

Fig. 9. The trajectories of the three agents in the musculoskeletal model reaching task.

TABLE III
THE ERRORS OF THE THREE AGENTS IN THE MUSCULOSKELETAL MODEL REACHING TASK

Target	EH-Agent	O-Agent	EO-Agent
0	0.211 ± 0.065	0.156 ± 0.098	0.054 ± 0.029
1	0.130 ± 0.012	0.112 ± 0.033	0.103 ± 0.029
2	0.068 ± 0.034	0.057 ± 0.032	0.041 ± 0.036
3	0.073 ± 0.022	0.104 ± 0.053	0.080 ± 0.023

icy coming into being, the population of rewards gets a smaller degree of dispersion, and the entropy gradually decreases as shown in Fig.10. Meanwhile, the value of normalized emotional valence remains growth with time, which indicates the effectiveness of this learning process. With the influence of valence, the neural system can adjust the learning parameters adaptively. As a decreasing function of valence, the learning rate and the variance of noise are large when the loss is low at the beginning, which makes the neural network explore more in action selection and learn with a high speed. Whereas, when the loss reaches a low residual value, the learning rate and noise become smaller for controlling the upper limb with great accuracy and stability. Additionally, by adaptively estimating the reward baseline, emotion can also improve the performance further, which embodies in two aspects. (1) At the initial stage, the predicted reward is more based on past experience so

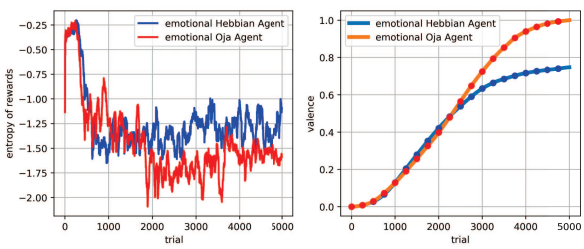

Fig. 10. The generation of the emotional valence. Left: the entropy of reward changes with the increase of trials. Right: the emotional valence changes with the increase of trials. Emotional Oja agent can obtain a higher value of valence than Hebbian one.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2018.2843563, IEEE Transactions on Cognitive and Developmental Systems

that current reward prediction error is usually large, and thus weights are adjusted with a large momentum. (2) However, at the latter phase, the predicted reward is more based on the nearest reward and changes dramatically, which enable the neural system to fine tune the weights better.

B. Robotic arm reaching task

For robotic learning problem, we investigate the effectiveness of our algorithm on a reaching task with a simulated Jaco arm. The physical Jaco with 6-DOF, developed by Kinova Robotics, is designed to be installed on a fixed surface or mobile platform. Here, according to official specifications and practical measurement, we build up the simulated arm in OpenSim. The goal of this task is to control the Jaco arm's end effector to reach to the ball at $(0.6, 0.8, 0.95)$ as shown in Fig.11. In the experiment, each trial is executed

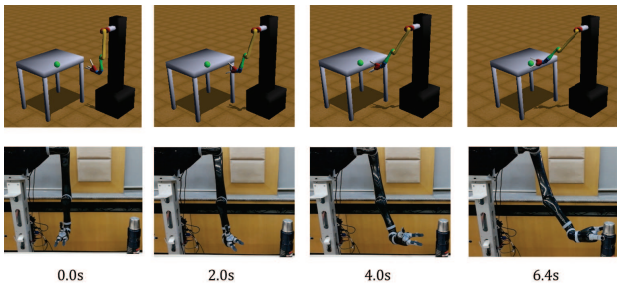

Fig. 11. Up: the scenario of the simulated robotic task. Down: the scenario of the real robotic implementation.

within 8.0s, where the population of neurons is warmed up across the first 1.6s and generates the control commands across later 6.4s. The control commands have 6 dimensions, which represents the desired angular velocity of each joint respectively. In practice, these joint velocity-control commands are published to physical system through ROS package for controlling the robotic arm in the real environment. A constant three-dimensional time series signal is fed into the network during the learning process. At each time step, the input vector is the coordinate of target point.

The parameter settings are close to the former experiment at all but the variance of the node perturbation is 0.4 by default. The population size is set to 400. We perform 2000 independent trials to reach to the target location, and take 20 tests to evaluate the performance. Obviously, the emotional Oja agent can also obtain a better performance with emotion modulation than two other agents in this robotic learning task (Fig.12). The mean and standard deviation of final position errors is shown in Table.IV.

A critical advantage of the proposed rule is that convergence time is independent of network size. Emotion-modulated Hebbian and Oja learning rules are used for the same task but with different network size. Loss values after different trials are plotted in Fig.13. The Hebbian approach loses performance clearly with the increase of network size, and fails to converge when the size is large. Because the search strategy of Hebbian rule is random so that it's hard to find a set of optimal weights in a huge space. However, Oja update rule can constrain the weights in a low-dimensional space no matter what the network size is. Thus, the performance of proposed method is effectively invariant with respect to the population size.

TABLE IV
THE ERRORS OF THE THREE AGENTS IN THE JACO REACHING TASK

	EH-Agent	O-Agent	EO-Agent
Position Error	0.167 ± 0.028	0.056 ± 0.012	0.020 ± 0.013

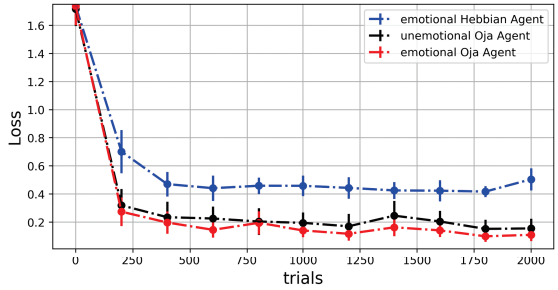

Fig. 12. The loss of the three agents in the robotic reaching task.

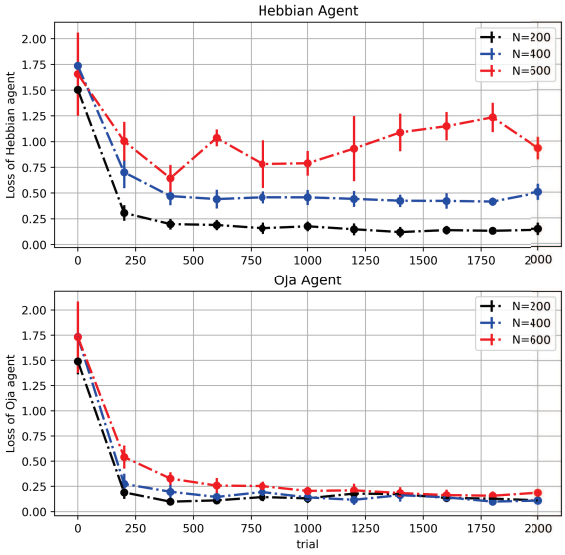

Fig. 13. The loss of the emotional Hebbian and Oja agents in the robotic arm reaching task. The size of population is set to 200, 400 and 600 respectively.

The changes of reward's entropy and normalized emotional valence are drawn in Fig.14. The entropy of rewards fluctuates in a relatively tight range, but overall, it tends to reduce. As uncertainty decreases, the emotional valence keeps growing, however the Oja agent gains the higher one than the Hebbian one.

Fig.15 shows the paths of end-effector movement in the test stage, where the ellipsoid denotes the range of the end point in each pattern of motion. The dashed lines represent the resulting optimal trajectories in the simulated environment, and the solid lines is actual trajectory of the real Jaco arm. The result demonstrates that the real robotic arm can perform the reaching task successfully with the learned policy in the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2018.2843563, IEEE Transactions on Cognitive and Developmental Systems

simulated environment. But there are still some errors, the reason of which might be that the controller's precision is not high enough. Additionally, Hebbian and Oja update rules make the recurrent neural network converge to different optimal value. The proposed algorithm allows the robotic arm to reach the goal with a higher accuracy.

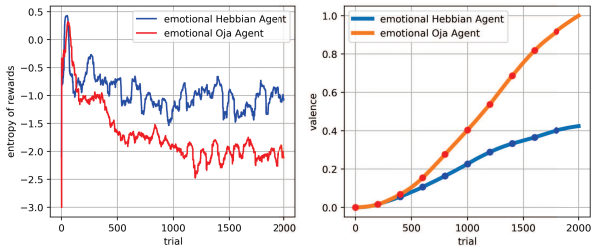

Fig. 14. The reward's entropy and emotional valence in the robotic arm reaching task.

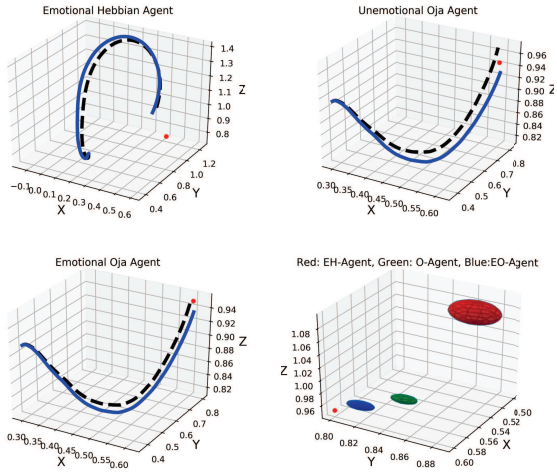

Fig. 15. 1,2,3: the trajectories of the three agents in the robotic arm reaching task. The dashed lines represent the resulting optimal trajectory in the simulated environment, and the solid lines are actual trajectories of the real Jaco arm. 4: the range of the last point in each pattern of motion.

V. Conclusion

In this paper, we present a novel emotion-modulated learning rule to train a recurrent neural network, which enables a neuro-muscle-skeleton and a robotic system to perform reaching tasks with higher accuracy and learning efficiency. Firstly, according to the neural mechanisms that motor cortex can generate complex spatiotemporal motor patterns through reward-based learning, a new method of reinforcement learning is proposed to train the recurrent neural network. Reward-modulated Hebbian and Oja update rules are applied to adjust internal recurrent weights with only delayed rewards at the end of each trial. Moreover, inspired by basic emotion theory on learning and decision making, we consider that emotion can dynamically modulate the parameters of learning or other decision making process through neuromodulatory system. More specifically, emotion can not only influence the reward prediction error by adjusting the reward prediction baseline, but also modulate the learning rate and the randomness in action selection. Hence, a computational model of emotion modulation is further developed to mediate the learning rule mentioned before. In the experimental part, we build a neuro-muscle-skeleton model of the human upper limb to perform reaching tasks through trial-and-reward learning, and use a simulated robotic arm Jaco to do reaching task as well. The results show that emotion-based learning methods can control the arm with a higher accuracy and a faster learning rate. Meanwhile, emotional Oja agent performs better than emotional Hebbian agent for ensuring the efficiency of learning.

References

[1] S. Shirafuji, S. Ikemoto, and K. Hosoda, "Development of a tendon-driven robotic finger for an anthropomorphic robotic hand," *The International Journal of Robotics Research*, vol. 33, no. 5, pp. 677–693, 2014.

[2] H. Qiao, C. Li, P. Yin, W. Wu, and Z.-Y. Liu, "Human-inspired motion model of upper-limb with fast response and learning ability – a promising direction for robot system and control," *Assembly Automation*, vol. 36, no. 1, pp. 97–107, 2016.

[3] A. L. Shoushtari, S. Mazzoleni, and P. Dario, "Bio-inspired kinematical control of redundant robotic manipulators," *Assembly Automation*, vol. 36, no. 2, pp. 200–215, 2016.

[4] D. G. Thelen and F. C. Anderson, "Using computed muscle control to generate forward dynamic simulations of human walking from experimental data." *Journal of Biomechanics*, vol. 39, no. 6, pp. 1107–1115, 2006.

[5] G. Rainer and E. K. Miller, "Effects of visual experience on the representation of objects in the prefrontal cortex," *Neuron*, vol. 27, no. 1, pp. 179–189, jul 2000.

[6] T. Klingberg, "Training and plasticity of working memory," *Trends in Cognitive Sciences*, vol. 14, no. 7, pp. 317–324, jul 2010.

[7] B. Jarosiewicz, S. M. Chase, G. W. Fraser, M. Velliste, R. E. Kass, and A. B. Schwartz, "Functional network reorganization during learning in a brain-computer interface paradigm," *Proceedings of the National Academy of Sciences*, vol. 105, no. 49, pp. 19 486–19 491, dec 2008.

[8] D. Sussillo and L. Abbott, "Generating coherent patterns of activity from chaotic neural networks," *Neuron*, vol. 63, no. 4, pp. 544–557, aug 2009.

[9] T. Miconi, "Biologically plausible learning in recurrent neural networks reproduces neural dynamics observed during cognitive tasks," *eLife*, vol. 6, feb 2017.

[10] X. Huang, W. Wu, P. Yin, and H. Qiao, "Improving learning efficiency of recurrent neural network through adjusting weights of all layers in a biologically-inspired framework," in *International Joint Conference on Neural Networks*, 2017, pp. 873–879.

[11] K. Rajan, C. D. Harvey, and D. W. Tank, "Recurrent network models of sequence generation and memory," *Neuron*, vol. 90, no. 1, pp. 128–142, apr 2016.

[12] H. F. Song, G. R. Yang, and X.-J. Wang, "Training excitatory-inhibitory recurrent neural networks for cognitive tasks: A simple and flexible framework," *PLOS Computational Biology*, vol. 12, no. 2, p. e1004792, feb 2016.

[13] G. M. Hoerzer, R. Legenstein, and W. Maass, "Emergence of complex computational structures from chaotic neural networks through reward-modulated hebbian learning," *Cerebral Cortex*, vol. 24, no. 3, pp. 677–690, nov 2012.

[14] A. F. Atiya and A. G. Parlos, "New results on recurrent network training: unifying the algorithms and accelerating convergence," *IEEE transactions on neural networks*, vol. 11, no. 3, pp. 697–709, 2000.

[15] J. J. Steil, "Backpropagation-decorrelation: Online recurrent learning with o (n) complexity," in *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*, vol. 2. IEEE, 2004, pp. 843–848.

[16] H. Jaeger, "Echo state network," *Scholarpedia*, vol. 2, no. 9, p. 2330, 2007.

[17] E. M. Izhikevich, "Solving the distal reward problem through linkage of STDP and dopamine signaling," *Cerebral Cortex*, vol. 17, no. 10, pp. 2443–2452, jan 2007.

[18] I. R. Fiete and H. S. Seung, "Gradient learning in spiking neural networks by dynamic perturbation of conductances," *Physical Review Letters*, vol. 97, no. 4, jul 2006.
[19] I. R. Fiete, M. S. Fee, and H. S. Seung, "Model of birdsong learning based on gradient estimation by dynamic perturbation of neural conductances," *Journal of Neurophysiology*, vol. 98, no. 4, pp. 2038–2057, aug 2007.
[20] T. M. Moerland, J. Broekens, and C. M. Jonker, "Emotion in reinforcement learning agents and robots: a survey," *Machine Learning*, no. 5, pp. 1–38, 2017.
[21] K. Doya, "Metalearning and neuromodulation," *Neural Networks*, vol. 15, no. 4-6, pp. 495–506, jun 2002.
[22] J. J. Gross, G. Sheppes, and H. L. Urry, "Emotion generation and emotion regulation: A distinction we should make (carefully)," *Cognition and emotion*, vol. 25, no. 5, pp. 765–781, 2011.
[23] K. N. Ochsner, R. R. Ray, B. Hughes, K. McRae, J. C. Cooper, J. Weber, J. D. Gabrieli, and J. J. Gross, "Bottom-up and top-down processes in emotion generation: common and distinct neural mechanisms," *Psychological science*, vol. 20, no. 11, pp. 1322–1331, 2009.
[24] A. R. Damasio, "Emotion and the human brain," *Annals of the New York Academy of Sciences*, vol. 935, no. 1, pp. 101–106, 2001.
[25] J. D. Cohen, "The vulcanization of the human brain: A neural perspective on interactions between cognition and emotion," *Journal of Economic Perspectives*, vol. 19, no. 4, pp. 3–24, nov 2005.
[26] E. T. Rolls, "Limbic systems for emotion and for memory, but no single limbic system," *Cortex*, vol. 62, pp. 119–157, jan 2015.
[27] E. A. Phelps, K. M. Lempert, and P. Sokol-Hessner, "Emotion and decision making: Multiple modulatory neural circuits," *Annual Review of Neuroscience*, vol. 37, no. 1, pp. 263–287, jul 2014.
[28] E. A. Phelps and J. E. LeDoux, "Contributions of the amygdala to emotion processing: From animal models to human behavior," *Neuron*, vol. 48, no. 2, pp. 175–187, oct 2005.
[29] H. Barbas, "Anatomic basis of cognitive-emotional interactions in the primate prefrontal cortex," *Neuroscience & Biobehavioral Reviews*, vol. 19, no. 3, pp. 499–510, 1995.
[30] R. S. Lazarus, *Emotion and adaptation*. Oxford University Press on Demand, 1991.
[31] S. Marsella, J. Gratch, P. Petta *et al.*, "Computational models of emotion," *A Blueprint for Affective Computing-A sourcebook and manual*, vol. 11, no. 1, pp. 21–46, 2010.
[32] H. L?vheim, "A new three-dimensional model for emotions and monoamine neurotransmitters," *Medical Hypotheses*, vol. 78, no. 2, pp. 341–348, feb 2012.
[33] J. LeDoux and J. R. Bemporad, "The emotional brain," *Journal of the American Academy of Psychoanalysis*, vol. 25, no. 3, pp. 525–528, 1997.
[34] J. E. LeDoux, "Emotion circuits in the brain," *Focus*, vol. 7, no. 2, pp. 274–274, 2009.
[35] J. E. LeDoux and J. M. Gorman, "A call to action: overcoming anxiety through active coping," *American Journal of Psychiatry*, vol. 158, no. 12, pp. 1953–1955, 2001.
[36] V. Van Veen and C. S. Carter, "The timing of action-monitoring processes in the anterior cingulate cortex," *Journal of cognitive neuroscience*, vol. 14, no. 4, pp. 593–602, 2002.
[37] B. A. Zavala, H. Tan, S. Little, K. Ashkan, M. Hariz, T. Foltynie, L. Zrinzo, K. A. Zaghloul, and P. Brown, "Midline frontal cortex low-frequency activity drives subthalamic nucleus oscillations during conflict," *Journal of Neuroscience*, vol. 34, no. 21, pp. 7322–7333, 2014.
[38] H. M. Bayer and P. W. Glimcher, "Midbrain dopamine neurons encode a quantitative reward prediction error signal," *Neuron*, vol. 47, no. 1, pp. 129–141, 2005.
[39] E. Tricomi and J. A. Fiez, "Feedback signals in the caudate reflect goal achievement on a declarative memory task," *Neuroimage*, vol. 41, no. 3, pp. 1154–1167, 2008.
[40] J. N. Reynolds and J. R. Wickens, "Dopamine-dependent plasticity of corticostriatal synapses," *Neural Networks*, vol. 15, no. 4, pp. 507–521, 2002.
[41] K. Doya, "Metalearning, neuromodulation, and emotion," in *Conference on Affective Minds*, vol. 46, 2000, p. 47.
[42] A. Etkin, C. Bchel, and J. J. Gross, "The neural bases of emotion regulation," *Nature Reviews Neuroscience*, vol. 16, no. 11, pp. 693–700, oct 2015.
[43] W. Schultz, "Reward functions of the basal ganglia," *Journal of Neural Transmission*, vol. 123, no. 7, pp. 679–693, 2016.
[44] R. S. Sutton and A. G. Barto, "Toward a modern theory of adaptive networks: Expectation and prediction." *Psychological Review*, vol. 88, no. 2, pp. 135–170, 1981.
[45] M. Shidara, T. G. Aigner, and B. J. Richmond, "Neuronal signals in the monkey ventral striatum related to progress through a predictable series of trials," *Journal of neuroscience*, vol. 18, no. 7, pp. 2613–2625, 1998.
[46] P. S. Goldman-Rakic, C. Leranth, S. M. Williams, N. Mons, and M. Geffard, "Dopamine synaptic complex with pyramidal neurons in primate cerebral cortex," *Proceedings of the National Academy of Sciences*, vol. 86, no. 22, pp. 9015–9019, 1989.
[47] F. Dolcos, A. D. Iordan, and S. Dolcos, "Neural correlates of emotion–cognition interactions: A review of evidence from brain imaging investigations," *Journal of Cognitive Psychology*, vol. 23, no. 6, pp. 669–694, 2011.
[48] E. A. Phelps, K. M. Lempert, and P. Sokol-Hessner, "Emotion and decision making: multiple modulatory neural circuits," *Annual Review of Neuroscience*, vol. 37, pp. 263–287, 2014.
[49] D. O. Hebb, *The organization of behavior: A neuropsychological theory*. Psychology Press, 2005.
[50] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, no. 3-4, pp. 229–256, 1992.
[51] R. Legenstein, S. M. Chase, A. B. Schwartz, and W. Maass, "A reward-modulated hebbian learning rule can explain experimentally observed network reorganization in a brain control task," *Journal of Neuroscience*, vol. 30, no. 25, pp. 8400–8410, 2010.
[52] E. Oja, "Simplified neuron model as a principal component analyzer," *Journal of mathematical biology*, vol. 15, no. 3, pp. 267–273, 1982.
[53] ——, "Neural networks, principal components, and subspaces," *International journal of neural systems*, vol. 1, no. 01, pp. 61–68, 1989.
[54] N. Frémaux, H. Sprekeler, and W. Gerstner, "Functional requirements for reward-modulated spike-timing-dependent plasticity," *Journal of Neuroscience*, vol. 30, no. 40, pp. 13 326–13 337, 2010.
[55] M. E. Hasselmo, "Neuromodulation: acetylcholine and memory consolidation," *Trends in cognitive sciences*, vol. 3, no. 9, pp. 351–359, 1999.
[56] ——, "The role of acetylcholine in learning and memory," *Current opinion in neurobiology*, vol. 16, no. 6, pp. 710–715, 2006.
[57] M. Habib, M. Cassotti, S. Moutier, O. Houdé, and G. Borst, "Fear and anger have opposite effects on risk seeking in the gain frame," *Frontiers in psychology*, vol. 6, p. 253, 2015.
[58] J. N. Druckman and R. McDermott, "Emotion and the framing of risky choice," *Political behavior*, vol. 30, no. 3, pp. 297–321, 2008.
[59] K. R. Saul, X. Hu, C. M. Goehler, M. E. Vidt, M. Daly, A. Velisar, and W. M. Murray, "Benchmarking of dynamic simulation predictions in two software platforms using an upper limb musculoskeletal model," *Computer methods in biomechanics and biomedical engineering*, vol. 18, no. 13, pp. 1445–1458, 2015.
[60] D. G. Thelen *et al.*, "Adjustment of muscle mechanics model parameters to simulate dynamic contractions in older adults," *Transactions-American Society Of Mechanical Engineers Journal Of Biomechanical Engineering*, vol. 125, no. 1, pp. 70–77, 2003.

Xiao Huang received the B.S. degree in guidance, navigation and control from Central South University, Changsha, China, in 2015. He is currently pursuing the Ph.D. degree in control theory and control engineering with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His current research interests include brain-inspired computing, affective computing and machine learning.

Wei Wu received the B.Sc. degree in physics and M.Sc. degree in theoretical physics from Beijing Normal University, Beijing, China, in 2001 and 2004, respectively, and the Ph.D. degree in computational neuroscience from Johann Wolfgang Goethe University, Frankfurt, Germany, in 2008. He is currently an Associate Professor with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing.

Hong Qiao (SM'06) received the B.Eng. degree in hydraulics and control and the M.Eng. degree in robotics from Xian Jiaotong University, Xian, China, in 1986 and 1989, respectively, the M.Phil. degree in robotics control from the Industrial Control Center, University of Strathclyde, Strathclyde, U.K., in 1992, and the Ph.D. degree in robotics and artificial intelligence from De Montfort University, Leicester, U.K., in 1995. She was a University Research Fellow with De Montfort University from 1995 to 1997. She was a Research Assistant Professor with the Department of Manufacturing Engineering and Engineering Management, City University of Hong Kong, Hong Kong, from 1997 to 2000, where she was an Assistant Professor from 2000 to 2002. Since 2002, she has been a Lecturer with the School of Informatics, The University of Manchester, Manchester, U.K. She is currently a Professor with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China. She first proposed the concept of the attractive region in strategy investigation, which has successfully been applied by herself in robot assembly, robot grasping, and part recognition. Her current research interests include information-based strategy investigation, robotics and intelligent agents, animation, machine learning, and pattern recognition. Her work has been reported in the Advanced Manufacturing Alert (Wiley, 1999). Dr. Qiao is currently a member of the Administrative Committee of the IEEE Robotics and Automation Society (RAS), a member of the IEEE Medal for Environmental and Safety Technologies Committee, and a member of the Early Career Award Nomination Committee, the Most Active Technical Committee Award Nomination Committee, and the Industrial Activities Board for RAS. She is currently an Associate Editor of the IEEE TRANSACTIONS ON CYBERNETICS and the IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING. She is the Editor-in-Chief of the Assembly Automation.

Yidao Ji received the B.S. degree in Automation, M.S. degree in Control Engineering from University of Science and Technology Beijing, Beijing China, where he is currently working toward the Ph.D. degree in the Department of Control Science and Engineering. His research interests include unmanned aerial vehicle, complex system and humanoid robotics.