Research Article

Decentralized Reinforcement Learning Robust Optimal Tracking Control for Time Varying Constrained Reconfigurable Modular Robot Based on ACI and Q-Function

Bo Dong¹ and Yuanchun Li²

¹ Department of Communication Engineering, Jilin University, Changchun 130022, China

² Department of Control Engineering, Changchun University of Technology, Changchun 130012, China

Correspondence should be addressed to Yuanchun Li; yuanchun@jlu.edu.cn

Received 20 August 2013; Revised 13 November 2013; Accepted 13 November 2013

Academic Editor: M. Onder Efe

Copyright © 2013 B. Dong and Y. Li. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

A novel decentralized reinforcement learning robust optimal tracking control theory for time varying constrained reconfigurable modular robots based on action-critic-identifier (ACI) and state-action value function (Q-function) has been presented to solve the problem of the continuous time nonlinear optimal control policy for strongly coupled uncertainty robotic system. The dynamics of time varying constrained reconfigurable modular robot is described as a synthesis of interconnected subsystem, and continuous time state equation and Q-function have been designed in this paper. Combining with ACI and RBF network, the global uncertainty of the subsystem and the HJB (Hamilton-Jacobi-Bellman) equation have been estimated, where critic-NN and action-NN are used to approximate the optimal Q-function and the optimal control policy, and the identifier is adopted to identify the global uncertainty as well as RBF-NN which is used to update the weights of ACI-NN. On this basis, a novel decentralized robust optimal tracking controller of the subsystem is proposed, so that the subsystem can track the desired trajectory and the tracking error can converge to zero in a finite time. The stability of ACI and the robust optimal tracking controller are confirmed by Lyapunov theory. Finally, comparative simulation examples are presented to illustrate the effectiveness of the proposed ACI and decentralized control theory.

1. Introduction

Reconfigurable modular robot could transform its configuration depending on the different external situations and the requirements of the tasks. According to the concept of modular design and the decentralized control theory of the subsystem, reconfigurable modular robot can complete the task by changing its structure efficiently in different situations, without redesigning the control law. At the same time, reconfigurable modular robot possessed a good accuracy and flexibility.

Many scholars have studied the dynamics and the control method of reconfigurable modular robot. A novel VGSTA-ESO based decentralized ADRC control method for reconfigurable modular robot has been proposed in [1]. Through designing a high-precision VGSTA-ESO to estimate the dynamic model nonlinear terms and the interconnection terms of the subsystem, the joint trajectory tracking control

is implemented. Based on calculating torque, a robust fuzzy neural network controller is proposed in [2], which is used to solve the problem of model uncertainty in the process of model generating. In [3], it shows a decentralized adaptive fuzzy sliding mode control method of reconfigurable modular robot. The fuzzy logic system is used to approximate the unknown dynamics of the subsystem, and a sliding mode controller with an adaptive scheme is designed to avoid both the interconnection term and the fuzzy approximation error. A decentralized adaptive neural network control algorithm for reconfigurable manipulators is proposed in [4], where the neural networks are used to approximate the unknown dynamic functions and interconnections in the subsystem by using the adaptive algorithm. A new distributed control method is proposed in [5], which uses a decomposition algorithm to decompose the robot dynamic system into a number of dynamical systems, and the adaptive sliding mode controller is designed to offset the impact of model

uncertainty. An observer based decentralized adaptive fuzzy controller for reconfigurable manipulator is proposed in [6]; by designing the state observer, the adaptive fuzzy systems which are used to model the unknown dynamics of the subsystem and the interconnection term can be constructed by using the state estimations. Nevertheless, the requirement of the dynamics of the reconfigurable modular robot system is hard to be satisfied, either fully or even partially known. Moreover, because of there are strong coupling model uncertainties and interconnection terms of subsystems in the reconfigurable modular robot system, besides the processing load on the controller would be increased, the fact that the greater time delay and calculation error are easy to produce, so that it is too complicated to design the controllers by using the methods and algorithms above.

In recent years, as one of the most effective methods to solve the control problems for continuous time which strongly coupled with nonlinear system, the reinforcement learning algorithm has received extensive attention from scholars. Reinforcement learning [7, 8] is a kind of learning method mapping situations to actions, so as to maximize a numerical reward signal. Compared with supervised learning, reinforcement learning does not need to predict the mentor signal in various states, but learns in the process of interaction with the situation. Because of its adaptive optimization capability in the nonlinear model under the condition of uncertainty, reinforcement learning has a unique advantage to solve the problems of optimization strategies and the control method in terms of the complex models [9–11]. Zhang and his team presented an infinite time optimal tracking control scheme for discrete-time nonlinear system via the greedy HDP iteration algorithm [12–14]. According to the system transformation, the optimal tracking problem is transformed into an optimal regulation problem, and the greedy HDP iteration algorithm is introduced to deal with the regulation problem with the rigorous convergence analysis. Then, a data-driven robust approximate optimal tracking control is proposed by using the adaptive dynamic programming, as well as a data-driven model which is established by a recurrent neural network to reconstruct the unknown system dynamics by using available input-output data [15]. After this, they design a fuzzy critic estimator which is used to estimate the value function for nonlinear continuous-time system [16]. On this basis, a synchronization problem for an array of neural networks with hybrid coupling and interval time varying delay is concerned with an augmented Lyapunov-Krasovskii functional method [17]. The FRL scheme only using immediate reward and sufficient conditions is adopted to analyze the convergence of the optimal task performance. Bhasin presents a neural network control of a robot interacting with an uncertain viscoelastic environment [18]. In [18], a continuous controller is developed for a robot that moved in free space and then regulated the new coupled dynamic system to a desired setpoint. Khan et al. present an implementation of a model-free Q-learning based on the discrete model reference compliance controller for a humanoid robot arm [19], where reinforcement learning scheme uses a recently developed Q-learning scheme to develop an optimal policy online. Patchaikani et al. propose

an adaptive critic-based real-time redundancy resolution scheme for kinematic control of the redundant manipulator [20]. The kinematic control of the redundant manipulator is formulated as a discrete-time input affine system; and then, an optimal real-time redundancy resolution scheme is proposed. Although the research of reinforcement learning algorithm has been rapidly developed in recent years, there are still some deficiencies. For example, when there are multiple subsystems in the global system, the methods above could not handle the impacts of the interconnection terms between the subsystems. Meanwhile, the methods above are mostly used to solve the learning and optimization problems of the system itself, but when the external constraints exist in the system, these methods are no longer applicable. Therefore, it is an urgent issue to solve the problem of how to design a kind of robust reinforcement learning optimal control method in the case of the external constraints existing and multiple subsystems coupled in the system are the urgent problems to be solved.

In this paper, we presented a novel continuous time decentralized reinforcement learning robust optimal tracking control theory for the time varying constrained reconfigurable modular robot. Combining with ACI and RBF-NN, the critic-NN is used to estimate the optimal Q-function; the action-NN is proposed to approximate the optimal control policy; and then, the identifier is adopted to identify the global uncertainty, so that the HJB equation can be estimated and the estimation error is bounded and converged. Firstly, since the decentralized control method is adopted in this paper, which means that each joint subsystem owns a separate controller, thus the processing loads of the controllers are reduced greatly. Secondly, due to the fact that the time varying constraints can be compensated in the subsystems, therefore the proposed method in this paper is suitable for the reconfigurable modular robot in the time varying constrained outside environment. Thirdly, the proposed control method could compensate for the impacts of the model uncertainties and the interconnection terms on the system, so that it can make the subsystems track the desired trajectories and the tracking error can converge to zero in finite time.

2. Problem Formulation

Assume that the time varying external constraints for the end of reconfigurable modular robot is shown as

$$\Psi(q, t) = 0. \quad (1)$$

Here $q \in R^n$ is the vector of joint displacements. Function $\Psi : R^n \rightarrow R^m$, m is the dimension of the external limiting conditions. With the time varying constraints, the dynamics of a reconfigurable modular robot can be presented as follows:

$$M(q)\ddot{q} + C(q, \dot{q})\dot{q} + G(q) + F(q, \dot{q}) = u + J_\Psi^T(q, t)f. \quad (2)$$

$M(q) \in R^{n \times n}$ is the inertia matrix, $C(q, \dot{q}) \in R^n$ is the Coriolis and centripetal force, $G(q) \in R^n$ is the gravity term, $F(q, \dot{q})$ is the unmodeled dynamics including friction terms and external disturbances, $u \in R^n$ is the applied joint torque,

and $J_{\Psi}^T(q, t)f$ is the contact force generated by the contact of the end of the reconfigurable modular robot and external constraints.

After introducing m th constraints for the robot which works in the free space, because of the limitation of (1), the system lost m th degrees of freedom. Therefore, the degrees of freedom of the robot change from n to $(n - m)$, so that only $(n - m)$ independent joint displacements are needed to describe the system of restricted movement fully.

Define

$$q = \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} \quad q_1 \in R^{n-m}, q_2 \in R^m. \quad (3)$$

Putting the equation above into (1); then we can get that

$$\Psi(q_1, \Theta(q_1, t), t) = 0, \quad (4)$$

where

$$q_2 = \Theta(q_1, t). \quad (5)$$

Therefore, (3) can be described by joint displacement q_1 fully, shown as follows:

$$q = \begin{bmatrix} q_1 \\ \Theta(q_1, t) \end{bmatrix}. \quad (6)$$

The derivation of (6) is

$$\begin{aligned} \dot{q} &= \begin{bmatrix} \frac{\partial \Theta(q_1, t)}{\partial q_1} \dot{q}_1 + \frac{\partial \Theta(q_1, t)}{\partial t} \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial \Theta(q_1, t)}{\partial q_1} & 0 \\ \frac{\partial \Theta(q_1, t)}{\partial t} & I_m \end{bmatrix} \begin{bmatrix} \dot{q}_1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{\partial \Theta(q_1, t)}{\partial t} \end{bmatrix} \\ &= T\dot{\theta} + H. \end{aligned} \quad (7)$$

In (7),

$$\begin{aligned} T &= \begin{bmatrix} \frac{\partial \Theta(q_1, t)}{\partial q_1} & 0 \\ 0 & I_m \end{bmatrix} \in R^{n \times n}, \\ \theta &= \begin{bmatrix} \dot{q}_1 \\ 0 \end{bmatrix} \in R^n, \quad H = \begin{bmatrix} 0 \\ \frac{\partial \Theta(q_1, t)}{\partial t} \end{bmatrix} \in R^n. \end{aligned} \quad (8)$$

Therefore, the second derivation of q can be achieved easily as

$$\ddot{q} = T\ddot{\theta} + \dot{T}\dot{\theta} + \dot{H}. \quad (9)$$

Putting (7) and (9) into (2), we can get

$$\begin{aligned} u + J_{\Psi}^T(q, t)f &= M(q)(T\ddot{\theta} + \dot{T}\dot{\theta} + \dot{H}) \\ &\quad + C(q, \dot{q})(T\dot{\theta} + \dot{H}) + G(q) + F(q, \dot{q}). \end{aligned} \quad (10)$$

Define

$$E = \begin{bmatrix} I_{(n-m) \times (n-m)} \\ 0_{m \times (n-m)} \end{bmatrix} \in R^{n \times (n-m)}. \quad (11)$$

Therefore

$$\theta = \begin{bmatrix} \dot{q}_1 \\ 0 \end{bmatrix} = E\dot{q}_1 \quad (12)$$

So, (2) can be decomposed into the following form:

$$\begin{aligned} &\sum_{j=1}^n M_{ij}(q) \left[(TE\ddot{q}_1)_j + (\dot{T}E\dot{q}_1)_j + \dot{H}_j \right] \\ &\quad + \sum_{j=1}^n C_{ij}(q, \dot{q}) \left[(TE\dot{q}_1)_j + H_j \right] + \bar{G}_i(q) \\ &\quad + F_i(q_i, \dot{q}_i) - f_i = u_i. \end{aligned} \quad (13)$$

In the equation above, $(TE\ddot{q}_1)_j$, $(\dot{T}E\dot{q}_1)_j$, $(TE\dot{q}_1)_j$, and H_j are the j th element of $(TE\ddot{q}_1)$, $(\dot{T}E\dot{q}_1)$, $(TE\dot{q}_1)$ and H , respectively; $\bar{G}_i(q)$, $F_i(q_i, \dot{q}_i)$, and u_i are the i th element of $G(q)$, $F(q, \dot{q})$, and u ; f_i is the constraint force which suffered by the i th joint. $M_{ij}(q)$ and $C_{ij}(q, \dot{q})$ are the ij th element of $M(q)$ and $C(q, \dot{q})$, respectively. So, as shown in Figure 1, each subsystem dynamical model can be formulated in joint space as follows:

$$\begin{aligned} &M_i(q_i) \ddot{q}_i + C_i(q_i, \dot{q}_i) \dot{q}_i + G_i(q_i) + F_i(q_i, \dot{q}_i) + Z_i(q, \dot{q}, \ddot{q}) \\ &= u_i, \end{aligned} \quad (14)$$

$$\begin{aligned} Z_i(q, \dot{q}, \ddot{q}) &= \sum_{\substack{j=1 \\ j \neq i}}^n M_{ij}(q) \left[(TE\ddot{q}_1)_j + (\dot{T}E\dot{q}_1)_j + \dot{H}_j \right] \\ &\quad + M_{ii}(q) \left[(TE\ddot{q}_1)_i + (\dot{T}E\dot{q}_1)_i + \dot{H}_i \right] \\ &\quad - M_i(q_i) \ddot{q}_i + \sum_{\substack{j=1 \\ j \neq i}}^n C_{ij}(q, \dot{q}) \left[(TE\dot{q}_1)_j + H_j \right] \\ &\quad + C_{ii}(q, \dot{q}) \left[(TE\dot{q}_1)_i + H_i \right] \\ &\quad - C_i(q_i, \dot{q}_i) \dot{q}_i + [\bar{G}_i(q) - G_i(q_i)]. \end{aligned} \quad (15)$$

Let $x_i = [x_{i1}, x_{i2}]^T = [q_i, \dot{q}_i]^T$ for $i = 1, \dots, n$; then (10) can be presented by the following state equation:

$$S_i : \begin{cases} \dot{x}_{i1} = x_{i2} \\ \dot{x}_{i2} = -f(x_i, u_i) - h_i(q, \dot{q}, \ddot{q}) - f_i, \\ y_i = x_{i1}, \end{cases} \quad (16)$$

FIGURE 1: The architecture of the time varying constrained reconfigurable modular robot system.

where x_i is the state vector of subsystem S_i , y_i is the output of subsystem S_i , and $h_i(q, \dot{q}, \ddot{q})$ is the interconnection term of the subsystem. $f(x_i, u_i)$ and $h_i(q, \dot{q}, \ddot{q})$ can be defined as

$$\begin{aligned} f(x_i, u_i) &= M_i^{-1}(q_i) \begin{bmatrix} C_i(q_i, \dot{q}_i) \dot{q}_i + G_i(q_i) \\ + F_i(q_i, \dot{q}_i) - u_i \end{bmatrix} \\ h_i(q, \dot{q}, \ddot{q}) &= -M_i^{-1}(q_i) Z_i(q, \dot{q}, \ddot{q}). \end{aligned} \quad (17)$$

In response to the time varying constrained reconfigurable modular robot system, we need to design a decentralized robust optimal tracking control policy to make the subsystem track the desired trajectory, as well as the tracking error is converged and bounded.

3. Decentralized Reinforcement Learning Robust Optimal Tracking Control Based on ACI and Q-Function

Assumption 1. Desired trajectory y_{id} , $\dot{y}_{id}$, $\ddot{y}_{id}$ and input gain matrix $b_i(x_i)$ are bounded.

Then (16) can be transformed to the below. Consider

$$S_i : \begin{cases} \dot{x}_{i1} = x_{i2} \\ \dot{x}_{i2} = -[F(x_i, u_i) + h_i(q, \dot{q}, \ddot{q}) + f_i] + b_i(x_i) u_i, \\ y_i = x_{i1}, \end{cases} \quad (18)$$

where $F(x_i, u_i) = f(x_i, u_i) + b_i(x_i)u_i$.

Assumption 2. The interconnection terms are bounded, satisfying the following equation:

$$|h_i(q, \dot{q}, \ddot{q})| \leq \delta_{i0} + \sum_{j=1}^n \delta_{ij}(|s_{ij}|), \quad (19)$$

where $\delta_{i0} > 0$ is an unknown constant and $\delta_{ij}(|s_{ij}|) \geq 0$ is an unknown smooth Lipschitz function.

The trajectory tracking error of the joint subsystem c can be defined as

$$e_i(t) = x_i - y_{id}. \quad (20)$$

With regard to the continuous time state equation of the subsystem in (18) with the nonlinear function and interconnection terms, generally, the value function can be defined as

$$V_i^{u_i(e_i(t))}(e_i(t)) = \int_0^\infty r_i(e_i(t), u_i(e_i(t))) dt. \quad (21)$$

In order to facilitate the equation, we use e_i, u_i instead of $e_i(t), u_i(e_i(t))$. Since the trajectory y_{id} relies upon the control of the subsystem u_i for updating, in order to avoid the infinity results by using (21), we need to transform the value function into the following form:

$$V_i^{u_i}(e_i) = \int_0^\infty r_i(e_i(\tau), u_i(e_i(\tau))) d\tau, \quad t \leq \tau < \infty. \quad (22)$$

Thus, the optimal value function of the subsystem can be defined as follows:

$$V_i^*(e_i) = \min_{u_i} \int_0^\infty r_i(e_i(\tau), u_i(e_i(\tau))) d\tau. \quad (23)$$

Here $r_i(e_i, u_i)$ represents the reward function for the current state, shown as

$$r_i(e_i, u_i) = e_i^T Q_e e_i + u_i^T R u_i, \quad (24)$$

where Q_e and R are the positive definite matrixes.

Typically, recording the value of state-action pairs is more useful than recording the value of state only, since the state-action pairs are the predictions of the reward. Even if the reward value of a state is low, it does not mean that the value of state-action pairs is low too. If the state of the subsystem in a period time produces a higher reward, then it can still get a higher state-action value. Therefore, from a long term perspective, defining a suitable state-action value function (Q-function) can make actions produce more rewards [21, 22].

According to (23) and (24), the continuous-time optimal Q-function can be defined as

$$\begin{aligned} Q_i^*(e_i, a_i, u_i) &= r_i(e_i, a_i, u_i) + V_i^*(e_i, u_i) \\ &= r_i(e_i, a_i, u_i) \\ &\quad + \min_{u_i(e_i)} \int_0^\infty r_i(e_i(\tau), u_i(e_i(\tau))) d\tau. \end{aligned} \quad (25)$$

Assumption 3. The partial derivation of Q_i^* and $r_i(e_i, a_i, u_i)$ exist and they are continuous in the domain. According to (18) and (24), by using the control policy u_i , the optimal Q-function can satisfy the following Hamiltonian-Jacobi-Bellman equation [23]:

$$\begin{aligned} &HJB_i(e_i, u_i, \nabla Q_i^*) \\ &= \min_{u_i(e_i)} [r_i(e_i, a_i, u_i) \\ &\quad + \nabla Q_i^* (-F(e_i, u_i) - h_i(e, \dot{e}, \ddot{e}) - f_i + b_i(e_i) u_i)] \\ &= \min_{u_i(e_i)} [r_i(e_i, a_i, u_i) + \nabla Q_i^* \Phi_i(e_i, u_i)], \end{aligned} \quad (26)$$

where $\Phi_i(e_i, u_i) = -F(e_i, u_i) - h_i(e, \dot{e}, \ddot{e}) - f_i + b_i(e_i) u_i$ means the global uncertainty including the unknown dynamics of the subsystem and the interconnection term, and $\nabla Q_i^* = \partial Q_i^* / \partial e_i$ means the gradient of the optimal Q-function.

Lemma 4 (see [24]). *Considering dynamics of the subsystem of time varying constrained reconfigurable modular robot in (14), in order to ensure the minimum of the HJB equation (26) possessing the stationary point with respect to u_i , the optimal Q-function and the optimal control policy must satisfy the following conditions:*

- (1) $\partial HJB(e_i, u_i, \nabla Q_i) / \partial u_i = 0$
- (2) $\partial^2 HJB(e_i, u_i, \nabla Q_i) / (\partial u_i \times \partial u_i^T) \geq 0$.

The necessary conditions above lead us to the following results.

- (a) The bounded control policy can guarantee a local minimum of the HJB equation (26) and satisfy the constraints imposed on the control inputs.
- (b) The Hessian matrix is positive-definite, and the control police u_i can render the global minimum of the HJB equation.
- (c) If an optimal algorithm exists, it is unique.

According to Lemma 4, if the reward function is smooth, and the optimal control u_i^* is adopted, then the HJB equation satisfies the following equation:

$$\begin{aligned} HJB_i^*(e_i, u_i^*, \nabla Q_i^*) &= \min_{u_i^*(e_i)} [r_i(e_i, a_i, u_i^*) + \nabla Q_i^* \Phi_i(e_i, u_i^*)] \\ &= 0. \end{aligned} \quad (27)$$

And the optimal control can be expressed as follows:

$$\begin{aligned} u_i^*(e_i) &= \arg_{u_i^*} \min [HJB_i^*(e_i, u_i^*, \nabla Q_i^*)] \\ &= \frac{1}{2} R^{-1} b_i^T(e_i) \frac{\partial Q_i^*(e_i, a_i, u_i)^T}{\partial e_i}. \end{aligned} \quad (28)$$

If the optimal Q-function Q_i^* is continuous, derivable, and known, and the initial value $Q_i^*(0) = 0$, as well as the optimal control policy $u_i^*(e_i)$, and the global uncertainty of the subsystem $\Phi_i(e_i, u_i^*)$ is known, then the HJB equation in (27) is held and solvable. However, in the actual situation, Q_i^* is not derivable everywhere, and $u_i^*(e_i)$ and $\Phi_i(e_i, u_i^*)$ are unknown. Therefore, it is not feasible to solve the HJB equation by using average method. In this paper, we combine the action-critic identifier (ACI) with RBF neural network to estimate the optimal control policy, the optimal Q-function, and the global uncertainty of the subsystem. Action-NN is used to estimate $u_i^*(e_i)$ and is denoted as $\hat{u}_i(e_i)$; Q_i^* is estimated by critic-NN and expressed as $\hat{Q}_i$; then we use the robust neural network identifier to identify $\Phi_i(e_i, u_i^*)$, denoted as $\hat{\Phi}_i(e_i, u_i^*)$. The block diagram of the ACI architecture is shown in Figure 2.

The estimated HJB equation can be expressed as follows:

$$H\hat{J}B_i^*(e_i, \hat{u}_i, \nabla \hat{Q}_i) = \min_{u_i(e_i)} [r_i(e_i, a_i, \hat{u}_i) + \nabla \hat{Q}_i \hat{\Phi}_i(e_i, \hat{u}_i)]. \quad (29)$$

The identification error of the HJB equation above can be expressed as

$$\delta_{hi} = H\hat{J}B_i^*(e_i, \hat{u}_i, \nabla \hat{Q}_i) - HJB_i^*(e_i, u_i^*, \nabla Q_i^*). \quad (30)$$

A classic radial basis function of the neural network is proposed in [25], shown as (31):

$$N(x) = W^{*T} S(x) + \varepsilon(x), \quad (31)$$

Combining (35) with (38), we can get that

$$\dot{\widehat{W}}_{ic}(t) = -n_1 L_i \left(L_i^T \widehat{W}_{ic} + \delta_{hi} \right). \quad (40)$$

The update law of the weight for the action-NN is developed by a gradient descent algorithm, expressed as follows:

$$\begin{aligned} \dot{\widehat{W}}_{ia}(t) &= -n_2 S_{ia}(e_i) \\ &\times \left(\left(\widehat{W}_{ia}^T S_{ia}(e_i) + \frac{1}{2} R^{-1} b_i(e_i) \nabla S_{ic}(e_i)^T \widehat{W}_{ic} \right)^T \right). \end{aligned} \quad (41)$$

According to the estimation error of action-NN in (36), the optimal control $u_i^*(e_i)$ can minimize the optimal Q-function, and we can get the following equation:

$$\begin{aligned} W_{ia}^T S_{ia}(e_i) + \frac{1}{2} R^{-1} b_i(e_i) \nabla S_{ic}(e_i)^T W_{ic} \\ + \frac{1}{2} R^{-1} b_i(e_i) \nabla \varepsilon_{ic}(e_i) + \varepsilon_{ia}(e_i) = 0. \end{aligned} \quad (42)$$

Putting (41) into (42), we can get that

$$\begin{aligned} \dot{\widehat{W}}_{ia} \\ = -n_2 S_{ia}(e_i) \left(\begin{aligned} &\widehat{W}_{ia}^T S_{ia}(e_i) + \frac{1}{2} R^{-1} b_i(e_i) \nabla S_{ic}(e_i)^T \widehat{W}_{ic} \\ &-\frac{1}{2} R^{-1} b_i(e_i) \nabla \varepsilon_{ic}(e_i) - \varepsilon_{ia}(e_i) \end{aligned} \right). \end{aligned} \quad (43)$$

After using critic-NN and action-NN to estimate $\widehat{Q}_i$ and $\widehat{u}_i(e_i)$, we need to design a kind of robust RBF-NN identifier to identify the nonlinear uncertainties of the subsystem. Here $\Phi_i(e_i, \widehat{u}_i)$ can be expressed as follows:

$$\Phi_i(e_i, \widehat{u}_i) = \dot{e}_{iF} = W_{iF}^T \kappa(\Lambda_{iF}^T e_{iF}) + \varepsilon_{iF}(e_{iF}) + b_i(e_i) \widehat{u}_i, \quad (44)$$

where $\kappa(\cdot)$ means the basic function of neural network, and W_{iF}, Λ_{iF} indicate the unknown ideal neural network weights. Equation (44) can be identified by using robust RBF-NN identifier, so we can get

$$\widehat{\Phi}_i(\widehat{e}_i, \widehat{u}_i) = \dot{\widehat{e}}_{iF} = \widehat{W}_{iF}^T \widehat{\kappa}_{iF} + b_i(e_i) \widehat{u}_i + \mu_i. \quad (45)$$

Here $\widehat{\kappa}_{iF}$ indicates the estimated value of the basic function of the neural network W_{iF}, Λ_{iF} are expressed as the estimated value of neural network. $\mu_i \in \mathbb{R}$ means the feedback error term, shown as follows [26]:

$$\begin{aligned} \mu_i &= k(e_{iF}(t) - \widehat{e}_{iF}(t)) - k(e_{iF}(0) - \widehat{e}_{iF}(0)) + \vartheta \\ &= k(\widetilde{e}_{iF}(t) - \widetilde{e}_{iF}(0)) + \vartheta \\ \dot{\vartheta} &= (k\alpha + \gamma)e_{iF} + \beta_1 \text{sat}(e_{iF}), \end{aligned} \quad (46)$$

where k, α, β_1 , and γ are the positive control gain constants and $\text{sat}(\cdot)$ is a saturation function. Therefore, the state estimation error of the identifier-NN can be expressed as follows:

$$\begin{aligned} \dot{\widetilde{e}}_{iF} &= \dot{e}_{iF} - \dot{\widehat{e}}_{iF} \\ &= W_{iF}^T \kappa_{iF} - \widehat{W}_{iF}^T \widehat{\kappa}_{iF} + \varepsilon_{iF}(e_{iF}) - \mu_i. \end{aligned} \quad (47)$$

A filtered identification error is defined as follows:

$$E_i = \dot{\widetilde{e}}_{iF} + \alpha \widetilde{e}_{iF}. \quad (48)$$

The derivation of the equation above is shown as

$$\begin{aligned} \dot{E}_i &= W_{iF}^T \dot{\kappa}_{iF} \Lambda_{iF}^T \dot{e}_{iF} - \widehat{W}_{iF}^T \dot{\widehat{\kappa}}_{iF} \widehat{\Lambda}_{iF}^T \dot{\widehat{e}}_{iF} + \alpha \dot{\widetilde{e}}_{iF} - k E_i - \gamma \widetilde{e}_{iF} \\ &\quad - \dot{\widehat{W}}_{iF}^T \widehat{\kappa}_{iF} + \dot{e}_{iF}(e_{iF}) - \beta_1 \text{sat}(\widetilde{e}_{iF}) - \widehat{W}_{iF}^T \dot{\widehat{\kappa}}_{iF} \widehat{\Lambda}_{iF}^T \dot{\widehat{e}}_{iF}. \end{aligned} \quad (49)$$

Here, the weight W_{iF}, Λ_{iF} of the identification-NN can be updated by

$$\begin{aligned} \dot{\widehat{W}}_{iF} &= \text{proj}(\Gamma_{iWF} \dot{\widehat{\kappa}}_{iF} \widehat{\Lambda}_{iF}^T \dot{\widehat{e}}_{iF} \widetilde{e}_{iF}^T) \\ \dot{\widehat{\Lambda}}_{iF} &= \text{proj}(\Gamma_{i\Lambda F} \dot{\widehat{e}}_{iF} \widetilde{e}_{iF}^T \widehat{W}_{iF}^T \dot{\widehat{\kappa}}_{iF}), \end{aligned} \quad (50)$$

where $\Gamma_{iWF}, \Gamma_{i\Lambda F}$ are positive constant adaptation gain matrices. In order to analyze the convergence of the filtered identification error, $\widehat{W}_{iF}^T \dot{\widehat{\kappa}}_{iF} \widehat{\Lambda}_{iF}^T \dot{\widehat{e}}_{iF}$ can be divided into the following form:

$$\begin{aligned} \widehat{W}_{iF}^T \dot{\widehat{\kappa}}_{iF} \widehat{\Lambda}_{iF}^T \dot{\widehat{e}}_{iF} \\ = \frac{1}{2} \dot{\widehat{\kappa}}_{iF} \dot{\widehat{e}}_{iF} \left[\left(\Lambda_{iF}^T - \widehat{\Lambda}_{iF}^T \right) \left(W_{iF}^T - \widehat{W}_{iF}^T \right) \right] \\ = \frac{1}{2} \dot{\widehat{\kappa}}_{iF} \dot{\widehat{e}}_{iF} \left[\begin{aligned} &\widehat{W}_{iF}^T \left(\Lambda_{iF}^T - \widehat{\Lambda}_{iF}^T \right) + \left(W_{iF}^T - \widehat{W}_{iF}^T \right) \widehat{\Lambda}_{iF}^T \\ &- W_{iF}^T \left(\Lambda_{iF}^T - \widehat{\Lambda}_{iF}^T \right) - \left(W_{iF}^T - \widehat{W}_{iF}^T \right) \Lambda_{iF}^T \end{aligned} \right] \\ = \frac{1}{2} \dot{\widehat{\kappa}}_{iF} \dot{\widehat{e}}_{iF} \left[\begin{aligned} &\widehat{W}_{iF}^T \left(\Lambda_{iF}^T - \widehat{\Lambda}_{iF}^T \right) + \left(W_{iF}^T - \widehat{W}_{iF}^T \right) \widehat{\Lambda}_{iF}^T \\ &- \frac{1}{2} \dot{\widehat{\kappa}}_{iF} \dot{\widehat{e}}_{iF} \left[W_{iF}^T \left(\Lambda_{iF}^T - \widehat{\Lambda}_{iF}^T \right) + \left(W_{iF}^T - \widehat{W}_{iF}^T \right) \Lambda_{iF}^T \right] \end{aligned} \right] \\ = \frac{1}{2} W_{iF}^T \dot{\widehat{\kappa}}_{iF} \widehat{\Lambda}_{iF}^T \dot{\widehat{e}}_{iF} + \frac{1}{2} \widehat{W}_{iF}^T \dot{\widehat{\kappa}}_{iF} \Lambda_{iF}^T \dot{\widehat{e}}_{iF} \\ - \frac{1}{2} W_{iF}^T \dot{\widehat{\kappa}}_{iF} \widehat{\Lambda}_{iF}^T \dot{\widehat{e}}_{iF} - \frac{1}{2} \widehat{W}_{iF}^T \dot{\widehat{\kappa}}_{iF} \Lambda_{iF}^T \dot{\widehat{e}}_{iF} \\ + \frac{1}{2} \widehat{W}_{iF}^T \dot{\widehat{\kappa}}_{iF} \widehat{\Lambda}_{iF}^T \dot{\widehat{e}}_{iF} + \frac{1}{2} \widehat{W}_{iF}^T \dot{\widehat{\kappa}}_{iF} \Lambda_{iF}^T \dot{\widehat{e}}_{iF}, \end{aligned} \quad (51)$$

where $\widehat{W}_{iF}^T = W_{iF}^T - \widehat{W}_{iF}^T, \widehat{\Lambda}_{iF}^T = \Lambda_{iF}^T - \widehat{\Lambda}_{iF}^T$. Putting (51) into (49), then (49) can be reduced to the following form:

$$\dot{E}_i = P_{F1} + P_{F2} + P_{F3} - k E_i - \gamma \widetilde{e}_{iF} - \beta_1 \text{sat}(\widetilde{e}_{iF}). \quad (52)$$

Among the equations above, $P_{F1} + P_{F2} + P_{F3}$ can be expressed, respectively, as follows:

$$P_{F1} = \frac{1}{2} W_{iF}^T \dot{\hat{\kappa}}_{iF} \hat{\Lambda}_{iF}^T \dot{\hat{e}}_{iF} + \frac{1}{2} \widehat{W}_{iF}^T \dot{\hat{\kappa}}_{iF} \Lambda_{iF}^T \dot{\hat{e}}_{iF} - \widehat{W}_{iF}^T \dot{\hat{\kappa}}_{iF} \hat{\Lambda}_{iF}^T \hat{e}_{iF} + \alpha \dot{\hat{e}}_{iF} - \widehat{W}_{iF}^T \hat{\kappa}_{iF} \quad (53)$$

$$P_{F2} = -\frac{1}{2} W_{iF}^T \dot{\hat{\kappa}}_{iF} \hat{\Lambda}_{iF}^T \dot{\hat{e}}_{iF} - \frac{1}{2} \widehat{W}_{iF}^T \dot{\hat{\kappa}}_{iF} \Lambda_{iF}^T \dot{\hat{e}}_{iF} + W_{iF}^T \dot{\hat{\kappa}}_{iF} \Lambda_{iF}^T \dot{\hat{e}}_{iF} + \dot{\hat{e}}_{iF} (e_{iF}) \quad (54)$$

$$P_{F3} = \frac{1}{2} \widehat{W}_{iF}^T \dot{\hat{\kappa}}_{iF} \hat{\Lambda}_{iF}^T \dot{\hat{e}}_{iF} + \frac{1}{2} \widehat{W}_{iF}^T \dot{\hat{\kappa}}_{iF} \Lambda_{iF}^T \dot{\hat{e}}_{iF}. \quad (55)$$

According to Assumption 1, (48), and (50), the upper bounds of P_{F1} , P_{F2} , P_{F3} are shown as

$$\begin{aligned} \|P_{F1}\| &\leq J_1 (\|\varphi_i(\tilde{e}_{iF}^T, E_i^T)\|) \|\varphi_i(\tilde{e}_{iF}^T, E_i^T)\| \\ \|P_{F2}\| &\leq \varsigma_1 \\ \|P_{F3}\| &\leq \varsigma_2. \end{aligned} \quad (56)$$

Combining (53) and (54) with (55), then we can get that:

$$\|\dot{P}_{F2} + \dot{P}_{F3}\| \leq \varsigma_3 + \varsigma_4 J_2 (\|\varphi_i(\tilde{e}_{iF}^T, E_i^T)\|) \|\varphi_i(\tilde{e}_{iF}^T, E_i^T)\|, \quad (57)$$

where $\varphi_i(\tilde{e}_{iF}^T, E_i^T) = [\tilde{e}_{iF}^T \ E_i^T]^T$ and $J_i(\cdot)$ is a global invertible nondecreasing function. ς_i , ($i = 1, 2, 3, 4$) are computable positive constants.

Theorem 5. *Considering dynamics of the subsystem of time varying constrained reconfigurable modular robot in (14) and the state equation (18), if the designed identifier and the corresponding weight update laws are adopted, then the global uncertainty of the subsystem which depends explicitly on the error term can be identified, and the identification error is converged and bounded.*

Proof. Define the Lyapunov function as the follows:

$$V_{iL}(e_{iF}, E_i) = \frac{1}{2} E_i^T E_i + \frac{1}{2} \gamma \tilde{e}_{iF}^T \tilde{e}_{iF} + \bar{\omega}_i(t) + \phi_i(t). \quad (58)$$

In the equation above, $\bar{\omega}_i(t)$ and $\phi_i(t)$ can be expressed as follows:

$$\begin{aligned} \bar{\omega}_i(t) &= - \left[\begin{array}{c} E_i^T (P_{F2} - \beta_1 \text{sat}(\tilde{e}_{iF})) + \dot{\tilde{e}}_{iF}^T P_{F3} \\ -\beta_2 J_2 (\|\varphi_i(\tilde{e}_{iF}^T, E_i^T)\|) \|\varphi_i(\tilde{e}_{iF}^T, E_i^T)\| \|\tilde{e}_{iF}\| \end{array} \right], \\ \bar{\omega}_i(0) &= \beta_1 |\tilde{e}_{iF}(0)| - \tilde{e}_{iF}^T(0) (P_{F2}(0) + P_{F3}(0)) \end{aligned} \quad (59)$$

$$\phi_i(t) = \frac{1}{4} \alpha \left[\text{tr}(\widehat{W}_{iF}^T \Gamma_{iWF}^{-1} \widehat{W}_{iF}) + \text{tr}(\bar{\Lambda}_{iF}^T \Gamma_{i\Lambda F}^{-1} \bar{\Lambda}_{iF}) \right], \quad (60)$$

where $\text{tr}(\cdot)$ represents the trace of matrix. Defining $d = [E_i^T \ \tilde{e}_{iF}^T \ \bar{\omega}_i^{1/2} \ \phi_i^{1/2}]$, $\beta_1, \beta_2 \in \mathbb{R}$ are positive adaptation gains which are chosen to ensure $\bar{\omega}_i(t) \geq 0$, so we can get

$$U_1(d) \leq V_{iL}(e_{iF}, E_i) \leq U_2(d), \quad (61)$$

where

$$U_1(d) = \frac{1}{2} \min(1, \gamma) \|d\|^2 \quad (62)$$

$$U_2(d) = \max(1, \gamma) \|d\|^2.$$

The derivation of (58) is shown as follows:

$$\dot{V}_{iL}(e_{iF}, E_i) = \nabla V_{iL}^T K \left[\begin{array}{c} \dot{E}_i \ \dot{\tilde{e}}_{iF}^T \ \frac{1}{2} \bar{\omega}_i^{1/2} \dot{\bar{\omega}}_i \ \frac{1}{2} \phi_i^{1/2} \dot{\phi}_i \end{array} \right]^T, \quad (63)$$

where $K[\cdot]$ is expressed as a Filippov set [27].

So $\dot{V}_{iL}(e_{iF}, E_i)$ can be deformed as the following form:

$$\begin{aligned} \dot{V}_{iL}(e_{iF}, E_i) &= \left[E_i^T \ \gamma \tilde{e}_{iF}^T \ 2\bar{\omega}_i^{1/2} \ 2\phi_i^{1/2} \right] K \left[\begin{array}{c} \dot{E}_i \ \dot{\tilde{e}}_{iF}^T \ \frac{1}{2} \bar{\omega}_i^{1/2} \dot{\bar{\omega}}_i \ \frac{1}{2} \phi_i^{1/2} \dot{\phi}_i \end{array} \right]^T \\ &\leq \gamma^T \left(\begin{array}{c} \frac{1}{2} W_{iF}^T \dot{\hat{\kappa}}_{iF} \hat{\Lambda}_{iF}^T \dot{\hat{e}}_{iF} + \frac{1}{2} \widehat{W}_{iF}^T \dot{\hat{\kappa}}_{iF} \Lambda_{iF}^T \dot{\hat{e}}_{iF} - \widehat{W}_{iF}^T \dot{\hat{\kappa}}_{iF} \hat{\Lambda}_{iF}^T \hat{e}_{iF} \\ + \alpha \dot{\hat{e}}_{iF} - \frac{1}{2} W_{iF}^T \dot{\hat{\kappa}}_{iF} \hat{\Lambda}_{iF}^T \dot{\hat{e}}_{iF} - \frac{1}{2} \widehat{W}_{iF}^T \dot{\hat{\kappa}}_{iF} \Lambda_{iF}^T \dot{\hat{e}}_{iF} - \gamma \tilde{e}_{iF} \\ - \widehat{W}_{iF}^T \dot{\hat{\kappa}}_{iF} + W_{iF}^T \dot{\hat{\kappa}}_{iF} \Lambda_{iF}^T \dot{\hat{e}}_{iF} + \dot{\hat{e}}_{iF} (e_{iF}) + \frac{1}{2} \widehat{W}_{iF}^T \dot{\hat{\kappa}}_{iF} \hat{\Lambda}_{iF}^T \dot{\hat{e}}_{iF} \\ + \frac{1}{2} \widehat{W}_{iF}^T \dot{\hat{\kappa}}_{iF} \Lambda_{iF}^T \dot{\hat{e}}_{iF} - k E_i - \beta_1 K[\text{sat}(\tilde{e}_{iF})] \end{array} \right) \\ &\quad - E_i^T \left(\begin{array}{c} -\frac{1}{2} W_{iF}^T \dot{\hat{\kappa}}_{iF} \hat{\Lambda}_{iF}^T \dot{\hat{e}}_{iF} - \frac{1}{2} \widehat{W}_{iF}^T \dot{\hat{\kappa}}_{iF} \Lambda_{iF}^T \dot{\hat{e}}_{iF} \\ + W_{iF}^T \dot{\hat{\kappa}}_{iF} \Lambda_{iF}^T \dot{\hat{e}}_{iF} + \dot{\hat{e}}_{iF} (e_i) - \beta_1 K[\text{sat}(\tilde{e}_{iF})] \end{array} \right) \\ &\quad - \tilde{e}_{iF}^T \frac{1}{2} \left(\widehat{W}_{iF}^T \dot{\hat{\kappa}}_{iF} \hat{\Lambda}_{iF}^T \dot{\hat{e}}_{iF} + \frac{1}{2} \widehat{W}_{iF}^T \dot{\hat{\kappa}}_{iF} \Lambda_{iF}^T \dot{\hat{e}}_{iF} \right) \\ &\quad + \gamma \tilde{e}_{iF}^T (E_i - \alpha \tilde{e}_{iF}) \\ &\quad + \beta_2 J_2 (\|\varphi_i(\tilde{e}_{iF}^T, E_i^T)\|) \|\varphi_i(\tilde{e}_{iF}^T, E_i^T)\| \|\tilde{e}_{iF}\| \\ &\quad - \frac{1}{2} \alpha \text{tr}(\widehat{W}_{iF}^T \Gamma_{iWF}^{-1} \widehat{W}_{iF}) - \frac{1}{2} \alpha \text{tr}(\bar{\Lambda}_{iF}^T \Gamma_{i\Lambda F}^{-1} \bar{\Lambda}_{iF}). \end{aligned} \quad (64)$$

Put (53), (54), and (55) into (64); then we can get:

$$\begin{aligned} \dot{V}_{iL}(e_{iF}, E_i) &= E_i^T (P_{F1} + P_{F2} + P_{F3} - \beta_1 K[\text{sat}(\tilde{e}_{iF})] - k E_i - \gamma \tilde{e}_{iF}) \\ &\quad + \gamma \tilde{e}_{iF}^T (E_i - \alpha \tilde{e}_{iF}) \\ &\quad - E_i^T (P_{F2} - \beta_1 K[\text{sat}(\tilde{e}_{iF})]) \\ &\quad - \tilde{e}_{iF}^T P_{F3} + \beta_2 J_2 (\|\varphi_i(\tilde{e}_{iF}^T, E_i^T)\|) \|\varphi_i(\tilde{e}_{iF}^T, E_i^T)\| \|\tilde{e}_{iF}\| \\ &\quad - \frac{1}{2} \alpha \text{tr}(\widehat{W}_{iF}^T \Gamma_{iWF}^{-1} \widehat{W}_{iF}) \\ &\quad - \frac{1}{2} \alpha \text{tr}(\bar{\Lambda}_{iF}^T \Gamma_{i\Lambda F}^{-1} \bar{\Lambda}_{iF}) \end{aligned}$$

$$\begin{aligned}
&= -\alpha \gamma \tilde{e}_{iF}^T \tilde{e}_{iF} + \left(E_i^T - \dot{\tilde{e}}_{iF}^T \right) P_{F3} \frac{J_1 \left(\left\| \varphi_i \left(\tilde{e}_{iF}^T, E_i^T \right) \right\| \right)^2}{4k_3} \\
&\quad - \frac{1}{2} \alpha \operatorname{tr} \left(\tilde{W}_{iF}^T \dot{\tilde{K}}_{iF} \tilde{\Lambda}_{iF}^T \dot{\tilde{e}}_{iF} \tilde{e}_{iF}^T \right) - \frac{1}{2} \alpha \operatorname{tr} \left(\tilde{\Lambda}_{iF}^T \dot{\tilde{e}}_{iF} \tilde{e}_{iF}^T \tilde{W}_{iF}^T \dot{\tilde{K}}_{iF} \right) \\
&\leq -k_1 \left\| \tilde{e}_{iF} \right\|^2 - k_2 \left\| E_i \right\|^2 \\
&\quad + \frac{J_1 \left(\left\| \varphi_i \left(\tilde{e}_{iF}^T, E_i^T \right) \right\| \right)^2}{4k_3} \left\| \varphi_i \left(\tilde{e}_{iF}^T, E_i^T \right) \right\|^2 \\
&\quad + \frac{\beta_2^2 J_2 \left(\left\| \varphi_i \left(\tilde{e}_{iF}^T, E_i^T \right) \right\| \right)^2}{4\alpha k_4} \left\| \varphi_i \left(\tilde{e}_{iF}^T, E_i^T \right) \right\|^2,
\end{aligned} \tag{65}$$

where $k_{\min} = \min\{k_1, k_2\}$, $\xi = \min\{k_3, \alpha k_4 / \beta_2^2\}$, and $J(\left\| \varphi_i(\tilde{e}_{iF}^T, E_i^T) \right\|)^2 = J_1(\left\| \varphi_i(\tilde{e}_{iF}^T, E_i^T) \right\|)^2 + J_2(\left\| \varphi_i(\tilde{e}_{iF}^T, E_i^T) \right\|)^2$, so the following conclusion can be obtained:

$$\begin{aligned}
&\dot{V}_{iL}(e_{iF}, E_i) \\
&\leq -k_{\min} \left\| \varphi_i(\tilde{e}_{iF}^T, E_i^T) \right\|^2 + \frac{J(\left\| \varphi_i(\tilde{e}_{iF}^T, E_i^T) \right\|)^2 \left\| \varphi_i(\tilde{e}_{iF}^T, E_i^T) \right\|^2}{4\xi} \\
&\leq -c \left\| \varphi_i(\tilde{e}_{iF}^T, E_i^T) \right\|^2.
\end{aligned} \tag{66}$$

Therefore, for an arbitrary constant c , $-c \left\| \varphi_i(\tilde{e}_{iF}^T, E_i^T) \right\|^2$ is a negative semidefinite function which is defined in the adjustable interval D , expressed as follows:

$$D = \left\{ d(t) \mid \|d\| \leq J^{-1} \left(2\sqrt{k_{\min}\xi} \right) \right\}, \tag{67}$$

so that Lyapunov stability theory shows that the system is stable. In order to make the subsystem of time varying constrained reconfigurable modular robot tracking the desired trajectory progressively, in this paper, a novel decentralized reinforcement learning robust optimal tracking controller has been designed by using the robust term to compensate the neural network approximation errors. Design the robust control term as

$$u_{irb} = \frac{N_{rb} e_i}{e_i^T e_i + \zeta}. \tag{68}$$

In the equation above, $\zeta > 0$ is a constant. And N_{rb} can be expressed as

$$\begin{aligned}
N_{rb} \geq &\left[\frac{\delta_{hi}^2}{2n_1} + \frac{n_1 \left(-\varepsilon_{ia}(e_i) - R^{-1} b_i(e_i) \left(\nabla \varepsilon_{ic}(e_i) / 2 \right) \right)^2}{2n_2} \right. \\
&\left. + n_1 n_2 \left\| b_i(e_i) \right\|^2 \left(-\varepsilon_{ia}(e_i) - R^{-1} b_i(e_i) \frac{\nabla \varepsilon_{ic}(e_i)}{2} \right)^2 \right] \\
&\cdot \frac{(e_i^T e_i + \zeta)}{2n_1 n_2 \left\| b_i(e_i) \right\| e_i^T e_i}
\end{aligned}$$

$$\begin{aligned}
&\geq \left[n_1^2 \left(-\varepsilon_{ia}(e_i) - R^{-1} b_i(e_i) \frac{\nabla \varepsilon_{ic}(e_i)}{2} \right)^2 \right. \\
&\quad \left. + n_2 \delta_{hi}^2 + 2n_1^2 n_2 \left\| b_i(e_i) \right\|^2 \right. \\
&\quad \left. \cdot \left(-\varepsilon_{ia}(e_i) - R^{-1} b_i(e_i) \frac{\nabla \varepsilon_{ic}(e_i)}{2} \right)^2 \right] \\
&\quad \cdot \frac{(e_i^T e_i + \zeta)}{4n_1^2 n_2^2 \left\| b_i(e_i) \right\| e_i^T e_i}.
\end{aligned} \tag{69}$$

Therefore, the global control law can be designed as follows:

$$\begin{aligned}
u_{\text{mix}} &= u_i + u_{irb} \\
&= -\frac{1}{2} R^{-1} b_i^T(e_i) \dot{S}_{ia}(e_i)^T \tilde{W}_{ia} + \frac{N_{rb} e_i}{e_i^T e_i + \zeta}.
\end{aligned} \tag{70}$$

□

Theorem 6. Considering dynamics of the subsystem of time varying constrained reconfigurable modular robot in (14), if the system parameters conditions and the assumptions are held, the critic-NN, action-NN, and identifier are given by (33), (34), and (45), respectively, and the decentralized robust optimal tracking controller of the subsystem in (70) is adopted; then, the system is closed-loop stability and the desired trajectory can be tracked asymptotically by the actual output.

Proof. Design the Lyapunov function as follows:

$$\begin{aligned}
V_{iu}(e_i, u_{\text{mix}}) &= \frac{1}{2n_1} \operatorname{tr} \{ \tilde{W}_{ic}^T \tilde{W}_{ic} \} + \frac{n_1}{2n_2} \operatorname{tr} \{ \tilde{W}_{ia}^T \tilde{W}_{ia} \} \\
&\quad + n_1 n_2 \left[e_{iF}^T e_{iF} + \Xi \int_0^\infty r_i(e_i, u_{\text{mix}}) d\tau \right],
\end{aligned} \tag{71}$$

where $\Xi > 0$ is the undetermined parameter, $t \leq \tau < \infty$. The derivation of (71) is shown as follows:

$$\begin{aligned}
&\dot{V}_{iu}(e_i, u_{\text{mix}}) \\
&= \frac{1}{2n_1} \operatorname{tr} \{ \tilde{W}_{ic}^T \dot{\tilde{W}}_{ic} \} + \frac{n_1}{2n_2} \operatorname{tr} \{ \tilde{W}_{ia}^T \dot{\tilde{W}}_{ia} \} \\
&\quad + n_1 n_2 \left[e_{iF}^T \dot{e}_{iF} + \Xi r_i(e_i, u_{\text{mix}}) \right] \\
&= \frac{1}{2n_1} \operatorname{tr} \{ \tilde{W}_{ic}^T (-n_1 L_i (L_i^T \tilde{W}_{ic} + \delta_{hi})) \} \\
&\quad + \frac{n_1}{2n_2} \operatorname{tr} \\
&\quad \times \left\{ \tilde{W}_{ia}^T \left[-n_2 S_{ia}(e_i) \left(\begin{aligned} &\tilde{W}_{ia}^T S_{ia}(e_i) - \varepsilon_{ia}(e_i) \\ &+ \frac{1}{2} R^{-1} b_i(e_i) \nabla S_{ic}(e_i)^T \tilde{W}_{ic} \\ &- \frac{1}{2} R^{-1} b_i(e_i) \nabla \varepsilon_{ic}(e_i) \end{aligned} \right) \right] \right\} \\
&\quad + n_1 n_2 \left\{ e_{iF}^T \left(W_{iF}^T \kappa \left(\Lambda_{iF}^T e_i \right) + \varepsilon_{iF}(e_i) + b_i(e_i) \hat{u}_{\text{mix}} \right) \right. \\
&\quad \left. + \Xi (e_i^T Q_e e_i + u_{\text{mix}}^T R u_{\text{mix}}) \right\}
\end{aligned}$$

$$\begin{aligned}
&\leq -\left(L_{im}^2 - \frac{n_1}{2}L_{iM}^2\right)\|\widetilde{W}_{ic}\|^2 + \frac{1}{2n_1}\delta_{hi}^2 \\
&\quad - \left(n_1S_{iam}^2 - \frac{3}{4}n_1n_2S_{iaM}^2\right)\|\widetilde{W}_{ia}\|^2 \\
&\quad + \frac{n_1}{4n_2}\|R^{-1}\|^2\|b_i(e_i)\|^2\|\nabla S_{icM}^2\|\|\widetilde{W}_{ic}\|^2 \\
&\quad + \frac{n_1}{2n_2}\left(\frac{\varepsilon_{ia}(e_i) + R^{-1}b_i(e_i)\nabla\varepsilon_{ic}(e_i)}{2}\right)^T \\
&\quad \cdot \left(\frac{\varepsilon_{ia}(e_i) + R^{-1}b_i(e_i)\nabla\varepsilon_{ic}(e_i)}{2}\right) \\
&\quad + n_1n_2\|b_i(e_i)\|^2\varepsilon_{ia}(e_i)^T\varepsilon_{ia}(e_i) \\
&\quad + n_1n_2S_{iaM}^2\|b_i(e_i)\|^2\|\widetilde{W}_{ia}\|^2 \\
&\quad + n_1n_2\Xi\lambda_{\min}(Q_e)\|e_{iF}\|^2 \\
&\quad + n_1n_2\left(\|b_i(e_i)\|^2 - \Xi\lambda_{\min}(R)\right)\|u_{\text{mix}}\|^2.
\end{aligned} \tag{72}$$

If the following inequalities can satisfy

$$\begin{aligned}
&\lambda_{\min}\{Q_e\}\|e_{iF}\|_2^2 \leq e_{iF}^T Q_e e_{iF} \leq \lambda_{\max}\{Q_e\}\|e_{iF}\|_2^2 \\
&\lambda_{\min}\{R\}\|u_{\text{mix}}\|_2^2 \leq u_{\text{mix}}^T R u_{\text{mix}} \leq \lambda_{\max}\{R\}\|u_{\text{mix}}\|_2^2, \\
&\Xi > \frac{\|b_i(e_i)\|^2}{\lambda_{\min}\{R\}},
\end{aligned} \tag{73}$$

then, $\dot{V}_{iu}(e_i, u_{\text{mix}})$ can be further transformed as

$$\begin{aligned}
&\dot{V}_{iu}(e_i, u_{\text{mix}}) \\
&\leq -\left(L_{im}^2 - \frac{n_1}{2}L_{iM}^2 - \frac{n_1}{4n_2}\|R^{-1}\|^2\|b_i(e_i)\|^2\nabla S_{icM}^2\right)\|\widetilde{W}_{ic}\|^2 \\
&\quad - \left(n_1S_{iam}^2 - \frac{3}{4}n_1n_2S_{iaM}^2 - n_1n_2S_{iaM}^2\|b_i(e_i)\|^2\right)\|\widetilde{W}_{ia}\|^2 \\
&\quad - n_1n_2\left(\|b_i(e_i)\|^2 + \Xi\lambda_{\min}(R)\right)\|u_{\text{mix}}\|^2 \\
&\quad - n_1n_2\Xi\lambda_{\min}(Q_e)\|e_{iF}\|^2 \\
&\leq -n_1n_2\Xi\lambda_{\min}(Q_e)\|e_{iF}\|^2.
\end{aligned} \tag{74}$$

Therefore, we can get the conclusion that $\dot{V}_{iu}(e_i, u_{\text{mix}}) < 0$. $\square$

4. Simulations

In order to verify the validity and convergence of the proposed decentralized reinforcement learning robust optimal tracking control method based on ACI and to study the convergence of the error by comparing the simulation result, in this paper, two different configurations of the time varying

external constrained reconfigurable modular robot have been applied, shown in Figures 3 and 4.

For the sake of the facilitation of the analysis of the configurations above, we can transform them into a form of analytic charts, which are shown in Figures 5 and 6, where L_1 , L_2 , and L_4 are the length of the links. L_3 is the distance between the time varying constraint joint and the base modular.

The time varying constraint can be defined as a kind of column which rotated about with a certain degree of freedom. The constraint equations of configuration A and configuration B are shown as follows:

$$\begin{aligned}
\Psi_A(q, t) &= L_1 \cos q_1 + L_2 \cos q_2 - [L_3 + L_4 \cot \alpha(t)] \\
\Psi_B(q, t) &= L_1 + L_2 \cos q_2 - [L_3 + L_4 \cot \alpha(t)].
\end{aligned} \tag{75}$$

In the equation above, the angle $\alpha(t)$ between the time varying constraint and the X-axis can be defined as follows:

$$\alpha(t) = 0.75\pi + 0.2 \sin \frac{t}{2}. \tag{76}$$

The initial positions of joint models are $q_1(0) = 2$, $q_2(0) = 2$ in configuration A and $q_1(0) = 2$, $q_2(0) = 2$ in configuration B. The initial velocities of joints are zeros. The dynamic model of configurations A and B is designed as follows:

$$\begin{aligned}
M_A(q) &= \begin{bmatrix} 0.36 \cos(q_2) + 0.6066 & 0.18 \cos(q_2) + 0.1233 \\ 0.18 \cos(q_2) + 0.1233 & 0.1233 \end{bmatrix} \\
M_B(q) &= \begin{bmatrix} 0.17 - 0.1166 \cos^2(q_2) & -0.06 \cos(q_2) \\ -0.06 \cos(q_2) & 0.1233 \end{bmatrix} \\
C_A(q, \dot{q}) &= \begin{bmatrix} -0.36 \sin(q_2) \dot{q}_2 & -0.18 \sin(q_2) \dot{q}_2 \\ 0.18 \sin(q_2) (\dot{q}_1 - \dot{q}_2) & 0.18 \sin(q_2) \dot{q}_1 \end{bmatrix} \\
C_B(q, \dot{q}) &= \begin{bmatrix} 0.1166 \sin(2q_2) \dot{q}_2 & 0.06 \sin(q_2) \dot{q}_2 \\ 0.06 \sin(q_2) \dot{q}_2 & 0 \end{bmatrix} \\
G_A(q) &= \begin{bmatrix} -5.88 \sin(q_1 + q_2) - 17.64 \sin(q_1) \\ -5.88 \sin(q_1 + q_2) \end{bmatrix} \\
G_B(q) &= \begin{bmatrix} 0 \\ -5.88 \cos(q_2) \end{bmatrix} \\
F_A(q, \dot{q}) &= \begin{bmatrix} \dot{q}_1 + 10 \sin(3q_1) + 2 \operatorname{sgn}(\dot{q}_1) \\ 1.2\dot{q}_2 + 5 \sin(2q_2) + \operatorname{sgn}(\dot{q}_2) \end{bmatrix} \\
F_B(q, \dot{q}) &= \begin{bmatrix} 0 \\ 1.5\dot{q}_2 + \sin(q_2) + 1.2 \operatorname{sgn}(\dot{q}_2) \end{bmatrix}.
\end{aligned} \tag{77}$$

The desired trajectory of configurations A and B is shown as Configuration A:

$$y_{1d} = 0.5 \cos(t) + 0.2 \sin(3t)$$

$$y_{2d} = \Theta(y_{1d}, t)$$

$$= \arcsin \left[\frac{L_1 \sin(\alpha(t) - y_{1d}) - L_3 \sin(\alpha(t))}{L_2} \right] + \alpha(t), \tag{78}$$

FIGURE 3: Configuration A for simulation.

FIGURE 4: Configuration B for simulation.

Configuration B:

$$\begin{aligned}
 y_{1d} &= 0 \\
 y_{2d} &= \Theta(y_{1d}, t) \\
 &= \arcsin \left[\frac{L_1 \sin(\alpha(t)) - L_3 \sin(\alpha(t))}{L_2} \right] + \alpha(t).
 \end{aligned} \tag{79}$$

Due to the limit of dynamic external constraints, the variety of joint 1 in configuration B is zero.

In order to confirm that the adopted method can be applied in different configurations and to verify the tracking performance for the subsystem desired trajectory by using the ACI and RBF-NN based decentralized reinforcement learning robust optimal tracking control method, in this part, the comparative simulations which include the classic RBF neural network control method and ACI based decentralized reinforcement learning robust optimal tracking control method have been adopted, respectively.

From Figures 7, 8, 9, 10, 11, 12, 13, and 14, the joint tracking and error curves are shown by using classic RBF neural network [4] to compensate the effect of the dynamic nonlinear term and the interconnection term in the subsystem. Figures 7 to 10 show that the actual output trajectories take about 2 seconds to track the desired trajectories. This is due to the fact that the classical neural network method requires a longer training process and parameter adaptive process. The error curves in Figures 11–14 show that the joint subsystem

FIGURE 5: The analytic chart of configuration A.

FIGURE 6: The analytic chart of configuration B.

constraints cannot be well compensated by adopting the classical neural control method for the reconfigurable modular robot when the time varying constraints exist. When the joint output variables turn larger, the reconfigurable modular system cannot exhibit a good robustness, and the tracking errors are larger than before.

Figures 15, 16, 17, 18, 19, 20, 21, and 22 showed the joint tracking and error curves by adopting ACI to identify the optimal Q-function, optimal control policy, and global dynamic nonlinear terms of the HJB equation in the subsystem. Figures 15 to 18 show that the actual output trajectories of the subsystems can track the desired trajectories in 0.5 seconds by using the proposed robust reinforcement learning optimal control method. This is due to the excellent identifying ability of ACI which can identify the uncertainties contained in the subsystems in a short time. The error curves in Figures 19–22 show that the tracking error is very small and it can converge to zero in a short time. Besides, when the proposed decentralized control method is adopted for the joint subsystems, the robustness is manifested.

Figures 23 and 24 showed the 3D-tip trajectory curves by using ACI algorithm. These two figures show that the proposed decentralized control method can fully satisfy the accessibility requirement of the reconfigurable modular robot. Besides, the singular displacements of the joint subsystems and the end-effector would not appear. The parameters defined in ACI are shown in Table 1.

FIGURE 7: Trajectory tracking curve of configuration A joint 1 with RBF neural network.

FIGURE 9: Trajectory tracking curve of configuration B joint 1 with RBF neural network.

FIGURE 8: Trajectory tracking curve of configuration A joint 2 with RBF neural network.

FIGURE 10: Trajectory tracking curve of configuration B joint 2 with RBF neural network.

TABLE 1: Parameter list of action-critic-identifier.

k	α	v	η_{a1}	η_{a2}	η_c	β_1	β_2	γ
800	300	0.005	10	50	20	0.2	2	0.5

The simulation results show that, compared with the situation of using classic RBF neural network controller, the decentralized robust optimal tracking control method based on ACI and reinforcement learning can be applied into different configurations of the time varying constrained reconfigurable modular robot. And the joint variables can track the desired trajectory within a very short time in different configurations, and the fluctuation of the error convergence range is minimal.

FIGURE 11: Tracking error curve of configuration A joint 1 with RBF neural network.

FIGURE 12: Tracking error curve of configuration A joint 2 with RBF neural network.

FIGURE 15: Trajectory tracking curve of configuration A joint 1 with ACI based reinforcement learning.

FIGURE 13: Tracking error curve of configuration B joint 1 with RBF neural network.

FIGURE 16: Trajectory tracking curve of configuration A joint 2 with ACI based reinforcement learning.

FIGURE 14: Tracking error curve of configuration B joint 2 with RBF neural network.

5. Conclusions and Future Work

In this paper, combining ACI with RBF neural network, a novel decentralized reinforcement learning robust optimal tracking control theory has been proposed for time varying constrained reconfigurable modular robots. Moreover, this theory is used to solve the problem of the continuous time nonlinear optimal control policy for strongly coupled uncertainty robotic system. Firstly, we build the model of subsystem with the time varying external constraints and described the global robot system as a synthesis of interconnected subsystem. Secondly, ACI is used to estimate the HJB equation and the global uncertainty, where a continuous-time optimal Q -function is adopted to take the place of traditional optimal value function, the optimal Q -function

FIGURE 17: Trajectory tracking curve of configuration B joint 1 with ACI based reinforcement learning.

FIGURE 18: Trajectory tracking curve of configuration B joint 2 with ACI based reinforcement learning.

and the optimal control policy are approximated by critic-NN and action-NN, and the global uncertainty is identified by the identifier. Thirdly, we design a novel decentralized robust optimal tracking controller, so the desired trajectory can be tracked, and the tracking error could converge to zero in finite time. On this basis, two kinds of Lyapunov functions are designed to confirm the stability of ACI and the subsystem. Finally, in order to confirm the superiority of the proposed control theory, the comparative simulation examples have been presented combining two different configurations of the robot.

In the future, more complex configuration for time varying external constrained reconfigurable modular robot can be included. Therefore, the decentralized controller with higher control and error precision will be considered.

FIGURE 19: Tracking error curve of configuration A joint 1 with ACI based reinforcement learning.

FIGURE 20: Tracking error curve of configuration A joint 2 with ACI based reinforcement learning.

FIGURE 21: Tracking error curve of configuration B joint 1 with ACI based reinforcement learning.

FIGURE 22: Tracking error curve of configuration B joint 2 with ACI based reinforcement learning.

FIGURE 23: 3D-tip trajectory curve of configuration A with ACI.

FIGURE 24: 3D-tip trajectory curve of configuration B with ACI.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work is financially supported by the National Natural Science Foundation of China (Grant nos. 61374051 and 60974010), the Scientific Technological Development Plan Project in Jilin Province of China (Grant no. 20110705). The first author is funded by China Scholarship Council.

References

- [1] Y. Li and B. Dong, "Decentralized ADRC control for reconfigurable manipulators based on VGSTA-ESO of sliding mode," *INFORMATION*, vol. 15, no. 6, pp. 2453–2466, 2012.
- [2] Y. Li, M.-C. Zhu, and Y.-C. Li, "Simulation on robust neurofuzzy compensator for reconfigurable manipulator motion control," *Journal of System Simulation*, vol. 19, no. 22, pp. 5169–5174, 2007.
- [3] M.-C. Zhu and Y.-C. Li, "Decentralized adaptive sliding mode control for reconfigurable manipulators using fuzzy logic," *Journal of Jilin University (Engineering and Technology Edition)*, vol. 39, no. 1, pp. 170–176, 2009.
- [4] L. Zhu and Y. Li, "Decentralized adaptive neural network control for reconfigurable manipulators," in *Proceedings of the Chinese Control and Decision Conference (CCDC '10)*, pp. 1760–1765, Xuzhou, China, May 2010.
- [5] M. Zhu, Y. Li, and Y. Li, "A new distributed control scheme of modular and reconfigurable robots," in *Proceedings of the IEEE International Conference on Mechatronics and Automation (ICMA '07)*, pp. 2622–2627, Harbin, China, August 2007.
- [6] M. C. Zhu, Y. Li, and Y. C. Li, "Observer-based decentralized adaptive fuzzy control for reconfigurable manipulator," *Control and Decision*, vol. 24, no. 3, pp. 429–434, 2009.
- [7] S. Richard and A. G. Barto, *Reinforcement Learning: An Introduction*, The MIT Press, London, UK, 1998.
- [8] F. L. Lewis and D. Liu, *Reinforcement Learning and Approximate Dynamic Programming For Feedback Control*, Wiley-IEEE Press, New York, NY, USA, 2012.
- [9] Y.-K. Xu and X.-R. Cao, "Lebesgue-sampling-based optimal control problems with time aggregation," *IEEE Transactions on Automatic Control*, vol. 56, no. 5, pp. 1097–1109, 2011.
- [10] F. L. Lewis and D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," *IEEE Circuits and Systems Magazine*, vol. 9, no. 3, pp. 32–50, 2009.
- [11] X. Xu, H.-G. He, and D. Hu, "Efficient reinforcement learning using recursive least-squares methods," *Journal of Artificial Intelligence Research*, vol. 16, pp. 259–292, 2002.
- [12] H. Zhang, Q. Wei, and Y. Luo, "A novel infinite-time optimal tracking control scheme for a class of discrete-time nonlinear systems via the greedy HDP iteration algorithm," *IEEE Transactions on Systems, Man, and Cybernetics B*, vol. 38, no. 4, pp. 937–942, 2008.
- [13] H. Zhang, R. Song, Q. Wei, and T. Zhang, "Optimal tracking control for a class of nonlinear discrete-time systems with time delays based on heuristic dynamic programming," *IEEE Transactions on Neural Networks*, vol. 22, no. 12, pp. 1851–1862, 2011.

- [14] H. Zhang, X. Xie, and S. Tong, "Homogenous polynomially parameter-dependent H_∞ filter designs of discrete-time fuzzy systems," *IEEE Transactions on Systems, Man, and Cybernetics B*, vol. 41, no. 5, pp. 1313–1322, 2011.
- [15] H. Zhang, L. Cui, X. Zhang, and Y. Luo, "Data-driven robust approximate optimal tracking control for unknown general nonlinear systems using adaptive dynamic programming method," *IEEE Transactions on Neural Networks*, vol. 22, no. 12, pp. 2226–2236, 2011.
- [16] J. Zhang, H. Zhang, Y. Luo, and H. Liang, "Optimal control design for nonlinear systems: adaptive dynamic programming based on fuzzy critic estimator," in *The International Joint Conference on Neural Network*, pp. 1–6, Brisbane, Australia, 2012.
- [17] H. Zhang, D. Gong, B. Chen, and Z. Liu, "Synchronization for coupled neural networks with interval delay: a novel augmented lyapunov-krasovskii functional method," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 1, pp. 58–70, 2013.
- [18] S. Bhasin, K. Dupree, P. M. Patre, and W. E. Dixon, "Neural network control of a robot interacting with an uncertain viscoelastic environment," *IEEE Transactions on Control Systems Technology*, vol. 19, no. 4, pp. 947–955, 2011.
- [19] S. G. Khan, G. Herrmann, F. Lewis, T. Pipe, and C. Melhuish, "A Q-learning based Cartesian model reference compliance controller implementation for a humanoid robot arm," in *Proceedings of the IEEE 5th International Conference on Robotics, Automation and Mechatronics (RAM '11)*, pp. 214–219, Qingdao, China, September 2011.
- [20] P. K. Patchaikani, L. Behera, and G. Prasad, "A single network adaptive critic-based redundancy resolution scheme for robot manipulators," *IEEE Transactions on Industrial Electronics*, vol. 59, no. 8, pp. 3241–3253, 2012.
- [21] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, no. 3–4, pp. 279–292, 1992.
- [22] C. J. C. H. Watkins, *Learning from delayed rewards [Ph.D. thesis]*, University of Cambridge, Cambridge, UK, 1989.
- [23] F. L. Lewis and V. L. Syrmos, *Optimal Control*, John Wiley & Sons, New York, NY, USA, 1995.
- [24] M. Sassano and A. Astolfi, "Dynamic approximate solutions of the HJ inequality and of the HJB equation for input-affine nonlinear systems," *IEEE Transactions on Automatic Control*, vol. 57, no. 10, pp. 2490–2503, 2012.
- [25] Y. X. Wu and C. Wang, "Deterministic learning based adaptive network control of robot in task space," *Acta Automatica Sinica*, vol. 39, no. 1, pp. 1–10, 2013.
- [26] P. M. Patre, W. MacKunis, K. Kaiser, and W. E. Dixon, "Asymptotic tracking for uncertain dynamic systems via a multilayer neural network feedforward and RISE feedback control structure," *IEEE Transactions on Automatic Control*, vol. 53, no. 9, pp. 2180–2185, 2008.
- [27] B. E. Paden and S. S. Sastry, "A calculus for computing Filippov's differential inclusion with application to the variable structure control of robot manipulators," *IEEE Transactions on Circuits and Systems*, vol. 34, no. 1, pp. 73–82, 1987.