IEEE Robotics and Automation Letters (RAL) paper presented at the
2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

A Multi-Channel Reinforcement Learning Framework for Robotic Mirror Therapy*

Jiajun Xu, Linsen Xu, Youfu Li, *Senior Member, IEEE*

Gaoxin Cheng, Jia Shi, Jinfu Liu and Shouqi Chen

Abstract— **In the paper, a robotic framework is proposed for hemiparesis rehabilitation. Mirror therapy is applied to transfer therapeutic training from the patient's function limb (FL) to the impaired limb (IL). The IL mimics the action prescribed by the FL with the assistance of the wearable robot, stimulating and strengthening the injured muscles through repetitive exercise. A master-slave robotic system is presented to implement the mirror therapy. Especially, the reinforcement learning is involved in the human-robot interaction control to enhance the rehabilitation efficacy and guarantee safety. Multi-channel sensed information, including the motion trajectory, muscle activation and the user's emotion, are incorporated in the learning algorithm. The muscle activation is expressed via the skin surface electromyography (EMG) signals, and the emotion is shown as the facial expression. The reinforcement learning approach is realized by the normalized advantage functions (NAF) algorithm. Then, a lower extremity rehabilitation robot with magnetorheological (MR) actuators is specially developed. The clinical experiments are carried out using the robot to verify the performance of the framework.**

I. Introduction

Hemiplegia has become increasingly common in recent years. The hemiplegic patient usually has the affected limbs on one unilateral side of the body, while the other side remains functional. To save cost of time and money due to expensive therapist labor, rehabilitation robots are developed. Mirror therapy proves an effective method for treatment of hemiparesis. A teleoperation architecture is built to transfer the motion from the physical therapist via the patient's functional limb (FL) to the impaired limb (IL) to complete the assist-as-needed (AAN) training [1]. However, the therapist is still included in the framework, which cannot actually realize saving of labor cost and in-home independent exercise. Though the bilateral impedance control [2] has been validated in the robotic mirror therapy, it is hard to clearly determine the impedance parameters and the adaption laws are uncertain. In this article, a novel controller based on reinforcement learning is proposed for more efficient recovery.

Reinforcement learning is one type of machine learning that maps states to actions with maximizing the numerical reward signal. A model-based framework is exploited to learn assistive strategy for wearable exoskeleton, and the user's muscular effort is taken into consideration while designing the cost function [3]. The integral reinforcement learning could be utilized to find the optimal parameters of the impedance model to adjust the robot's dynamics with respect to the operator skills [4]. As for this case in the paper, due to the uncertainty of the robot's and human's dynamic model and continuity of the manipulation trajectory, a model-free reinforcement learning algorithm that functions in the continuous action domain is required. Considering that, the normalized advantage functions (NAF) is employed [5].

Facial expression can be fused in the robotic therapy framework to detect the variance of patient's feeling in real time. For example, the hemiplegic patient may feel pain or uncomfortable during exercise and his/her IL cannot move voluntarily, and the response of the FL may be not instantaneous enough too, but this can be reflected by the facial expression timely. Hence, facial expression recognition (FER) is a vital method to show the subject's emotion. Convolutional neural network (CNN) is an effective model for massive image processing, and it has been adopted in deep learning for FER [6]. However, real-time FER becomes even more difficult because the facial expression turns to be dynamic in videos instead of static images, and especially the facial expression is more changeable as the training intensity varies. Encountered with this, 3D convolutional networks and long short-term memory (LSTM) are combined to extract temporal relations of consecutive frames in a video sequence.

The contribution of this paper is to construct a robotic framework for in-home hemiparesis rehabilitation. The control strategy based on reinforcement learning can efficiently improve the rehabilitation efficacy on the basis of safety guarantee. Besides, it has the potential to accommodate different patients with different movement abilities.

The remainder of the paper is organized as follows. The dynamics and transmission principle of the master-slave robotic system implementing mirror therapy is stated in Section II. The impedance control for the master robot is presented in Section III. The reinforcement learning control

*This work was supported by National Key Research and Development Plan #2017YFB1303200, Anhui Science and Technology Major Project #17030901034 and Key Research and Development Plan of Jiangsu Province #BE2017007-1.

Jiajun Xu is with University of Science and Technology of China and City University of Hong Kong, China (e-mail: jiajun@mail.ustc.edu.cn, jiajunxu2-c@my.cityu.edu.hk).

Linsen Xu is with Hefei Institutes of Physical Science, CAS, and the Key Laboratory of Biomimetic Sensing and Advanced Robot Technology, Anhui Province, No.801, Changwu Road, Changzhou, 213164, China (phone: +86 15995075378, e-mail: lsxu@iamt.ac.cn).

Youfu Li, Senior Member, IEEE, is with City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong, China (e-mail: meyfli@cityu.edu.hk).

Gaoxin Cheng, Jia Shi, Jinfu Liu and Shouqi Chen are with University of Science and Technology of China (email: {ba181681, sj1996, liujinfu, chenshouqi09}@mail.ustc.edu.cn).

for the slave robot is described in Section IV. Section V represents the specially designed hardware for the framework. Validation experiments are accomplished in Section VI. Finally, the paper is concluded in Section VII.

II. Problem Formulation

The bilateral master-slave robotic system is applied for mirror therapy for hemiplegia, where the master robot is attached to the FL and the slave robot is worn by the IL. The dynamic model of the robot in joint space is shown as

$$M_m(q_m)\ddot{q}_m + C_m(q_m, \dot{q}_m)\dot{q}_m + G_m(q_m) + f_m(\dot{q}_m) = \tau_m + \tau_{FL} \tag{1}$$

$$M_s(q_s)\ddot{q}_s + C_s(q_s, \dot{q}_s)\dot{q}_s + G_s(q_s) + f_s(\dot{q}_s) = \tau_s + \tau_{IL} \tag{2}$$

where $q_i \in \mathbb{R}^n (i = m, s$, m denotes the master robot and s denotes the slave robot, and n means the number of manipulator joints) is the position coordinates of the robotic joints, and accordingly $\dot{q}_i$ and $\ddot{q}_i$ represent the joint velocity and acceleration respectively. $M_i(q_i) \in \mathbb{R}^{n\times n}$ is the inertia matrix, $C_i(q_i, \dot{q}_i) \in \mathbb{R}^{n\times n}$ is the centripetal and Coriolis term, $G_i(q_i) \in \mathbb{R}^n$ is the gravity torque, and $f_i(\dot{q}_i)$ presents the friction torque. The parameter $\tau_i \in \mathbb{R}^n$ stands for the robotic torque generated by the actuator, $\tau_{FL} \in \mathbb{R}^n$ means the interaction torque exerted by the FL and $\tau_{IL} \in \mathbb{R}^n$ means the interaction torque exerted by the IL.

Some important properties of the dynamic model (1) and (2) are listed in the following.

Property 1: The matrix $\dot{M}_i(q_i) - 2C_i(q_i, \dot{q}_i)$ is skew-symmetric.

Property 2: The inverse matrix $M_i^{-1}(q)$ exists, and it is positive definite and bounded.

Property 3: The left side is linearly parameterized as $M_i(q_i)\ddot{q}_i + C_i(q_i, \dot{q}_i)\dot{q}_i + G_i(q_i) + f_i(\dot{q}_i) = W_i(\varphi_{i1}, \varphi_{i2}, q_i, \dot{q}_i)\theta_i$, where θ_i is a set of unknown constant parameters, and $W_i(\varphi_{i1}, \varphi_{i2}, q_i, \dot{q}_i)$ is a known dynamic regressor matrix.

In the bilateral master-slave robotic system, the position and velocity of the master robot is transmitted with time delay T_m to the slave side for the IL to follow. Also, the interaction torque between the slave robot and the IL is transmitted with time delay T_s to the master side. The transmission channels of the system are illustrated as Fig.1. Thus, we can obtain

$$\begin{aligned} q_{sd}(t) &= K_q q_m(t - T_m) \\ \dot{q}_{sd}(t) &= K_q \dot{q}_m(t - T_m) \\ \tau_{FLd}(t) &= K_\tau \tau_{IL}(t - T_s) \end{aligned} \tag{3}$$

where t is the current time, $q_{sd} \in \mathbb{R}^n$ is the desired position for the slave robot, and $\tau_{FLd} \in \mathbb{R}^n$ is the transmitted interaction torque to the FL side. The parameter $K_q = \mathrm{diag}(K_{q1}, \dots, K_{qn})$ and $K_\tau = \mathrm{diag}(K_{\tau 1}, \dots, K_{\tau n})$ means the mirroring matrix, accommodating for the mirroring effect between the FL and IL, and $K_{q1}, \dots, K_{qn}, K_{\tau 1}, \dots, K_{\tau n} = +1\ or - 1$ [1]. Take the lower extremity therapy for example, the element in the mirroring matrix should be +1 for the hip flexion/extension, knee flexion/extension and ankle dorsiflexion/plantar-flexion because of the same orientation, and it should be -1 for the hip abduction/adduction due to the opposite moving direction between two legs.

As for the robotic control strategy, the model reference adaptive impedance control and reinforcement learning control are employed for the master and slave robot respectively. The overall control diagram for the robot is depicted as Fig. 2, and the details will be explained in the following sections.

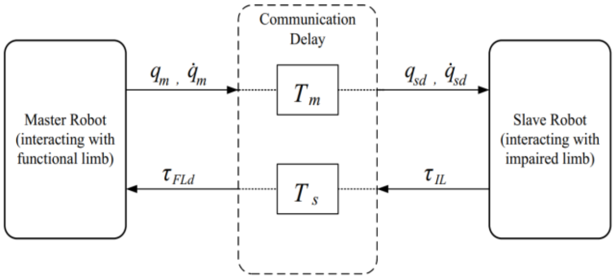

Communication Delay
q_m , $\dot{q}_m$
T_m
q_{sd} , $\dot{q}_{sd}$
Master Robot (interacting with functional limb)
Slave Robot (interacting with impaired limb)
τ_{FLd}
T_s
τ_{IL}

Figure 1. The transmission channels of the master-slave robotic system.

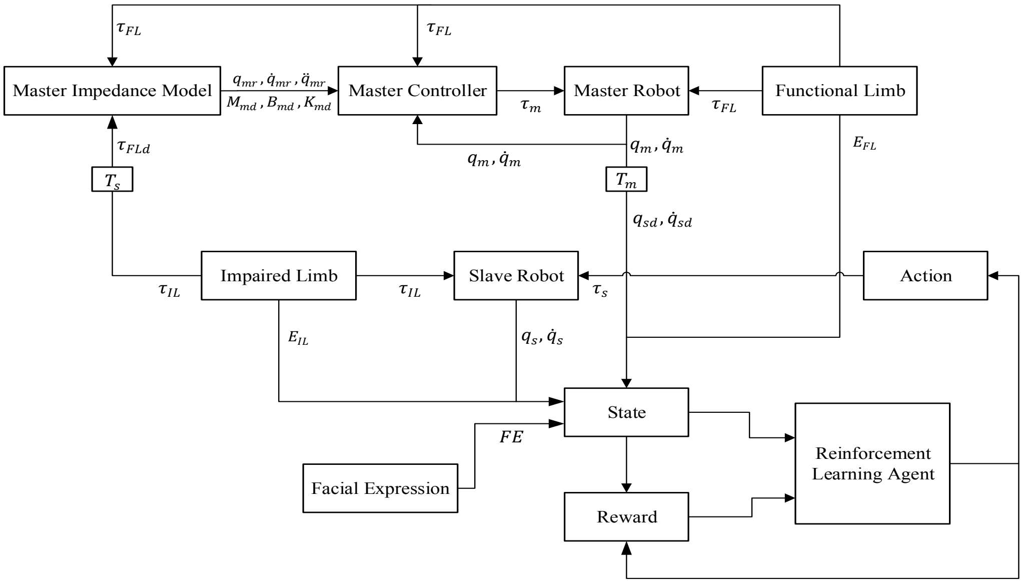


Figure 2. The overall control diagram.

III. Master Controller

In the bilateral robotic system, the model reference adaptive impedance control is used to design the master controller. First, the reference impedance model (4) receives the interaction torque from the master and slave side, generating the appropriate motion trajectory q_{mr}. For example, when the IL feels pain and uncomfortable during therapy and fails to move voluntarily, the FL can exert less muscle force (smaller τ_{FL}) to limit the motion range of the master robot (smaller q_{mr}), and the slave robot's movement is accordingly constrained. Otherwise, when the IL feels resistance or gets impairment and cannot complete the given task autonomously, the FL can increase effort (larger τ_{FL}) to expand the motion range for assistance (larger q_{mr}). Therefore, the subject can adjust the FL's force to modify the movement of the FL and IL for more comfort and proper exercise intensity. Then, the master controller is put forward to track the desired position q_{mr}. The reference impedance model of the master robot can be defined as

$$M_{md}\ddot{q}_{mr} + B_{md}\dot{q}_{mr} + K_{md}q_{mr} = \tau_{FL} - \tau_{FLd} \quad (4)$$

where $M_{md} \in \mathbb{R}^{n\times n}$, $B_m \in \mathbb{R}^{n\times n}$ and $K_{md} \in \mathbb{R}^{n\times n}$ refer to the desired inertia, damping and stiffness matrix of the master robot, $q_{mr} \in \mathbb{R}^n$ is the output of the impedance model, meaning the desired position, and accordingly $\dot{q}_{mr}$ and $\ddot{q}_{mr}$ represent the desired velocity and acceleration respectively.

The master controller is designed to ensure that the actual position asymptotically tracks the desired trajectory from the reference impedance model (4), i.e. $q_m \to q_{mr}$ and $\dot{q}_m \to \dot{q}_{mr}$. Thus, the sliding model variable s_m is given to minimize the tracking error as (5)

$$s_m = \dot{\tilde{q}}_m + \lambda_1\tilde{q}_m \quad (5)$$

where $\tilde{q}_m = q_m - q_{mr}$, and λ_1 is a positive constant. The reference velocity (6) is obtained to formulate the sliding mode error (5) as $s_m = \dot{q}_m - \dot{q}_r$.

$$\dot{q}_r = \dot{q}_{mr} - \lambda_1\tilde{q}_m \quad (6)$$

Since the acceleration of the master manipulator joints is required while designing the controller but its measurement is challenging, it can be estimated with satisfactory accuracy when the master robot mimics the impedance model (4). Therefore, the estimated acceleration for the master robot is

$$\hat{\ddot{q}}_m = -M_{md}^{-1}B_{md}\dot{q}_{mr} - M_{md}^{-1}K_{md}q_{mr} + M_{md}^{-1}(\tau_{FL} - \tau_{FLd}) - \lambda_1\dot{\tilde{q}}_m - \lambda_2 s_m \quad (7)$$

where λ_2 is a positive constant. The controller for the master robot in joint space is then designed as [7]

$$\tau_m = \widehat{M}_m(q_m)\hat{\ddot{q}}_m + \hat{C}_m(q_m,\dot{q}_m)\dot{q}_r + \hat{G}_m(q_m) + \hat{f}_m(\dot{q}_m) - \tau_{FL} \quad (8)$$

where the accent '^' denotes the estimated values. According to *Property 3*, the controller can be rewritten as

$$\tau_m = \widehat{M}_m(q_m)\varphi_{m1} + \hat{C}_m(q_m,\dot{q}_m)\varphi_{m2} + \hat{G}(q_m) + \hat{f}(\dot{q}_m) - \tau_{FL} \quad (9)$$

where $\varphi_{m1} = \hat{\ddot{q}}_m$, $\varphi_{m2} = \dot{q}_r$. So, the control input turns to

$$\tau_m = W_m(\varphi_{m1},\varphi_{m2},q_m,\dot{q}_m)\hat{\theta}_m - \tau_{FL} \quad (10)$$

where $\hat{\theta}_m$ is the estimation of θ_m. Substituting the controller (10) into the robot dynamics (1), one can obtain

$$M_m(q_m)\dot{s}_m = -\lambda_2 M_m(q_m)s_m - C_m(q_m,\dot{q}_m)s_m + W_m\tilde{\theta}_m \quad (11)$$

where $\tilde{\theta}_m = \hat{\theta}_m - \theta_m$.

In order to prove the tracking convergence ($q_m \to q_{mr}$), the Lyapunov function candidate is proposed as follows.

$$V(t) = \frac{1}{2}(s_m^T M_m(q_m)s_m + \tilde{\theta}_m^T\Gamma_m^{-1}\tilde{\theta}_m) \quad (12)$$

where Γ_m is a symmetric positive definite matrix. Then, the time derivative of $V(t)$ is derived by applying (11) and *Property 1*.

$$\dot{V}(t) = -\lambda_2 s_m^T M_m(q_m)s_m + s_m^T W_m\tilde{\theta}_m + \dot{\hat{\theta}}_m^T\Gamma_m^{-1}\tilde{\theta}_m (13)$$

The parameter adaption law is defined as

$$\dot{\hat{\theta}}_m = -\Gamma_m^T W_m^T s_m \quad (14)$$

So, the time derivative of $V(t)$ is simplified to

$$\dot{V}(t) = -\lambda_2 s_m^T M_m(q_m)s_m \quad (15)$$

Based on the positive definiteness of the inertia matrix M_m (seen in *Property 2*) and the adaption gain matrix Γ_m, the Lyapunov function candidate is positive definite, i.e. $V(t) \geq 0$, and its time derivative is negative definite, i.e. $\dot{V}(t) \leq 0$. It implies that the tracking error converges to zero, i.e. $q_m \to q_{mr}$ and $\dot{q}_m \to \dot{q}_{mr}$ as $t \to \infty$, and the system stability is satisfied.

IV. Slave Controller

Reinforcement learning is applied to design the slave controller, which aims to maximize the muscle activation of the IL and restrain the IL within the range of motion prescribed by the FL to ensure safety. The standard reinforcement learning setup consists of environment *E*, state *s*, action *a* and reward *r*. For each timestep *t,* the agent receives a state s_t that can be fully observable and then take an action a_t. A scalar reward r_t is then obtained for evaluation. The reward function is designed after the state s_t completes the action a_t as $r(s_t, a_t)$. Since the learning is a Markov decision process, the return from the state is thought as the discounted future reward $R_t = \sum_{i=t}^{T}\gamma^{i-t}r(s_i, a_i)$ with $\gamma \in [0,1]$ being the discounting factor. Due to the continuity of the robotic joint trajectory, we enable the basic reinforcement learning algorithm, Q-learning, in continuous action spaces with deep neural networks. The Q function $Q(s_t, a_t)$ is gotten as the expected return from s_t after taking the action a_t and following the current policy thereafter. To constitute the Q function, the Q network is built in the NAF algorithm, which receives the state *s* as input and outputs a value function term $V(s|\theta^V)$ and an advantage term $A(s,a|\theta^A)$. The advantage term is parameterized as a quadratic function of nonlinear features of the state as

$$A(s,a|\theta^A) = -\frac{1}{2}\big(a - \mu(s|\theta^\mu)\big)^T P(s|\theta^P)\big(a - \mu(s|\theta^\mu)\big)$$

where $\mu(s|\theta^\mu)$ is the action outputs from the neural network, and $P(s|\theta^P)$ is a state-dependent, positive-definite square matrix, which is defined as $P(s|\theta^P) = L(s|\theta^P)L(s|\theta^P)^T$, where $L(s|\theta^P)$ is a lower-triangular matrix whose entries come from a linear output layer of the neural network, with the diagonal terms exponentiated [5]. Summarily, the Q network outputs the value function $V(s|\theta^V)$, action $\mu(s|\theta^\mu)$ as well as the term to be transferred to $L(s|\theta^P)$, the latter two of which constitute the advantage. Then, the Q-function is obtained as

$$Q(s,a|\theta^Q) = A(s,a|\theta^A) + V(s|\theta^V)$$

where $\theta^Q = \{\theta^\mu, \theta^P, \theta^V\}$ is the parameter of the network. The learning objective is to maximize the Q function.

In the robotic rehabilitation framework, the robot's future state depends on its state, the human's state and the robot's action. The robot's state s_R includes the joint position and velocity of the master-slave robotic system, i.e.

$$s_R = [q_{m1}, \dot{q}_{m1} \dots q_{mn}, \dot{q}_{mn}, q_{s1}, \dot{q}_{s1} \dots q_{sn}, \dot{q}_{sn}]^T$$

where q_{mi} $(i = 1,2 \dots n)$ denotes the position of the ith joint of the master manipulator, and q_{si} $(i = 1,2 \dots n)$ denotes the position of the ith joint of the slave manipulator. Sequentially, $\dot{q}_{mi}$ and $\dot{q}_{si}$ refer to the ith joint velocity of the master and slave robot respectively. Notice that limitations of range of motion and joint torques are preset as the learning hyperparameters to guarantee the users' safety.

The patients' muscle activation is one main expression of the rehabilitation efficacy. In this project, the muscle activation is expressed in the form of the skin surface electromyography (EMG). The EMG signals are detected with the electrodes attached to the relevant skin surface. To remove noise and guarantee real-time response, the detected EMG signals should be filtered from 5 Hz to 500 Hz and then low pass filtered with the cut-off frequency less than 20 Hz, and the input latency should optimally be kept between 100 ms and 250 ms for pre-processing [8]. Basically, the higher the IL's EMG values get, the better the patient has recovered. Particularly, the proposed therapy framework focuses on the patients in the flaccid paralysis period whose upper motor neurons are not damaged badly. While the upper motor neurons are further damaged, spasticity may occur, and therefore the spasticity situation should be included in the human state. The spasticity can be found through detecting the abnormal EMG signals [9-10], and the robot's action requires to be ceased immediately with therapists and medicine involved.

Additionally, the user's emotion is included in the reinforcement learning framework and can control the exercise intensity in real time. The human emotion is shown as the facial expression, and it can be separated into seven classifications including happiness, surprise, anger, contempt, disgust, fear and sadness in the CK+ database [11]. We define the first two as the positive facial expression, and label the last five as the negative facial expression. When the positive facial expression is shown, the robot's action can be appropriately enhanced; otherwise, the exercise should be weakened and even stopped. The facial expression plays an additional role in the framework, and its effect is smaller than trajectory tracking error and muscle activation by modulating the reward function. But it is necessary especially when the patients feel painful during the therapy.

The selected network architecture in this paper is the 3D version of Inception-ResNet, which combines the advantages of residual connections of ResNet and Inception of GoogLeNet [12-13]. It extracts both spatial and temporal features of the sequences, and the overall network architecture is shown in Fig. 3. The input is the real-time facial expression videos, whose size is resized into 10×299×299×3 (10 frames, 299×299 frame size and 3 color channels). Then, the stem layer [13], 3D Inception-ResNet A, 3D Reduction A, 3D Inception-ResNet B, 3D Reduction B, 3D Inception-ResNet C, Average Pooling, Dropout, LSTM and a fully connected layer are followed. The recognized category of the facial expression is finally out. Among them, the Reduction layer (3D Reduction A and 3D Reduction B) is used to reduce the grid size. Particularly, the LSTM unit takes the enhanced feature map resulted from the 3D Inception-ResNet layer as an input by vectorizing the feature map on its sequence dimension, and the temporal information is extracted. The LSTM unit receives the input whose temporal length is six, and it contains 200 hidden units to fit the FER task. The LSTM unit is followed by a fully-connected layer associated with a softmax activation function. In addition, the facial landmarks are incorporated to help more accurate recognition by tracking the important facial components. The facial landmarks are incorporated by replacing the shortcut in residual unit on original ResNet with element-wise multiplication of facial landmarks and the input tensor of the residual unit, which is presented as the "Elem-Mul" module in Fig. 3. The FER experiments using the network and CK+ database have shown that the recognition rate is up to $93.21 \pm 2.32\%$, which is adequately high in comparison with the state of arts [12].

Hence, there are multiple channels for the reinforcement learning agent to receive the human states, including the EMG signals and facial expression, i.e.

$$s_H = [E_{FL1} \dots E_{FLk}, E_{IL1} \dots E_{ILk}, FE]^T$$

where E_{FLi} $(i = 1,2 \dots k$, k is the number of the measured muscles) means the EMG value of the ith muscle of the FL, and E_{ILi} is the EMG value of the ith muscle of the IL. The observed facial expression is saved as FE.

Therefore, the state of the human-robot interaction is

$$s = [s_R, s_H]^T$$

Besides, the robot's action is the torque generated by the joint actuators of the slave robot, such that

$$a = [\tau_{s1}, \dots, \tau_{sn}]^T$$

where τ_{si} $(i = 1,2 \dots n)$ is the actuator torque of the ith joint on the slave side.

The goal of the reinforcement learning control is to maximize the IL's muscle activation with the minimal trajectory tracking error between the master and slave manipulation. Facial expression is also adopted for additional assistance, and positive facial expression is encouraged while negative facial expression is discouraged. Different from using the absolute value of the EMG signal to design the reward function [3], the relative value is applied. The uniform rehabilitation strategy for different patients is to increase the IL's muscle activation to approach their own FL's, such that the proportion of the post-processed EMG level of the IL's certain muscle with respect to one of the FL, i.e. E_{ILi}/E_{FLi}, is taken into consideration while determining the reward function. Besides, the motion of the IL should be maintained along the trajectory of the FL to guarantee controllability and safety. So, the multi-channel reward function is designed as

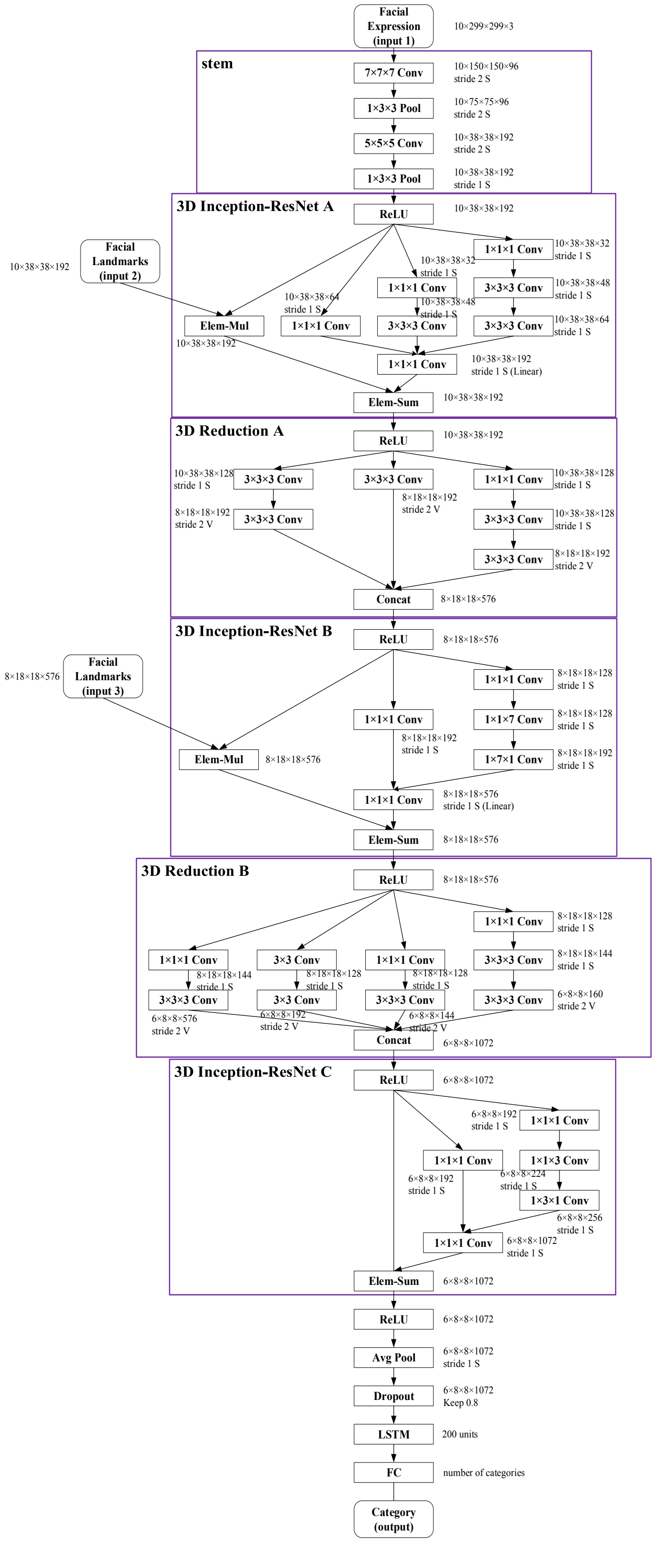


Figure 3. The network architecture for FER. The size of the output tensor is attached to each layer. The “V” and “S” marked layers mean Valid and Same paddings respectively. All of the convolution layers are followed by an ReLU activation function except the ones that are indicated as “Linear”.

$$r = -(q_s - q_{sd})^T \Lambda (q_s - q_{sd}) + \sum_{i=1}^{k} \gamma_{Ei} \frac{E_{ILi}}{E_{FLi}} + V_{FER} - \gamma_u \sum_{i=1}^{n} (u_{t-1}^i)^2 - \gamma_A \sum_{i=1}^{n} (\ddot{q}_s^i)^2 \qquad (16)$$

where Λ is the diagonal positive matrix that expresses the weight of the tracking error term in the reward function, the parameter γ_{Ei} is the constant balancing the contribution of the EMG signal of the *i*th measured muscle. The constant V_{FER} refers to the FER result. When the recognized result is the positive facial expression, $V_{FER} > 0$; otherwise, $V_{FER} < 0$. The absolute value of V_{FER} for negative facial expression should be larger than positive in that FER is mainly used to prevent injury during therapy. The action term is weighed by the constant γ_u, indicating that the robot support is not adequately required for stimulating muscle activation to fulfill the exercise and should be minimized. The parameter γ_A is the constant to balance the contribution of the acceleration term, and this term needs to be reduced to avoid abrupt motion and ensure the user’s safety. It is essential to determine each above-mentioned weight in the reward function (16), and we use the relative entropy inverse reinforcement learning algorithm to tune the weights [14].

The pseudocode of the learning is shown in Algorithm 1. The Q network is constructed in the NAF algorithm [5]. It contains a normalized Q network and a target Q network, and they share the same network architecture. The normalized Q network is randomly initialized with the weight θ^Q, and it is updated by sampling a minibatch from the replay buffer. The Q network is optimized by minimizing the loss function (18), which presents the Bellman error between the Q function under the normalized network and target network, and the target value is defined as (17). After that, the target network is updated by $\theta^{Q'} \leftarrow \varepsilon\theta^Q + (1-\varepsilon)\theta^{Q'}$ with ε denoting the learning rate, and then the learning proceeds into the next iteration.

Algorithm 1: Slave robot control algorithm based on NAF method

```
Initialize: discounting factor $\gamma$, learning rate $\varepsilon$, total episodes P,
            maximum timestep T, maximum iteration I
Initialize: normalized Q network $Q(s,a|\theta^Q)$ with the weight $\theta^Q$
Initialize: target Q network $Q'$ with the weight $\theta^{Q'} \leftarrow \theta^Q$
Initialize: replay buffer $D$
For episode=1: P do
    Initialize a random process $\eta$ for action exploration
    Receive the initial state $s_1$ by measuring the robotic joint positions,
    joint velocities and human skin surface EMGs through sensors
    For timestep=1: T do
        Select action $a_t = \mu(s_t|\theta^\mu) + \eta_t$ with $\eta_t$ being the current
         exploration noise
        Execute action $a_t$ and observe reward $r_t$, and then receive a new
         observation state $s_{t+1}$
        Store transition $(s_t, a_t, r_t, s_{t+1})$ in $D$
        For iteration=1: I do
        Sample a random minibatch of $N$ transitions $(s_i, a_i, r_i, s_{i+1})$ from $D$
        Set $y_i = r_i + \gamma V'(s_{i+1}|\theta^{Q'})$                              (17)
        Update $\theta^Q$ by minimizing the loss:
        $L = \frac{1}{N}\sum_i (y_i - Q(s_i, a_i|\theta^Q))^2$                         (18)
       Update the target networks: $\theta^{Q'} \leftarrow \varepsilon\theta^Q + (1-\varepsilon)\theta^{Q'}$
        end For
    end For
end For
```

V. Hardware

In order to realize the proposed therapy framework, a multi-mode lower extremity rehabilitation robot is designed and implemented. It has four degrees of freedom (DOFs) in single limb including hip abduction/adduction, hip flexion/extension, knee flexion/extension and ankle dorsiflexion/plantar-flexion, especially, among which the magnetorheological (MR) actuators are equipped at the first three joints [15-17]. The MR actuator can control its torque output by adjusting the input electric current. Compared with normal actuators, MR actuators can generate more flexible torque with little power consumption immediately, and accordingly safety is guaranteed in the rehabilitation therapy. The design scheme of the robot is shown in Fig. 4.

The robot has two working modes: robot-active mode and human-active mode, and the transition between the two modes is allowed [15]. In the robot-active mode, the MR actuator works as a clutch to transfer the motor torque to the robotic joint, and the robot leads the human leg to move. While in the human-active mode, the human leg dominantly guides the movement of the robotic exoskeleton, and the MR actuator functions as a brake to help the user conduct anti-resistance training to strengthen muscles. In this mirror therapy framework, the MR actuators at the master robot work in the human-active mode. This allows the transmitted interaction torque from the IL can produce impedance to control the motion of the master manipulator. Besides, the MR actuators equipped at the slave exoskeleton are tuned to the robot-active mode, and the IL is driven with assistance of the slave robot.

VI. Experiments

The proposed framework is verified with the above-mentioned robot. Two DOFs including hip flexion/extension (HFE) and knee flexion/extension (KFE) get involved in the experiments. Five hemiplegic patients (P1-P5) participate in the tests, and their information is listed in Table I, where the gender, age, years post-stroke, Fugl-Meyer Assessment (FMA) score before therapy and paretic side are included. The electrodes (ETS FreeEMG 300) are attached to the corresponding muscles for detecting the skin surface EMG signals. For HFE, six electrodes are attached to six muscles: gluteus maximus (GM), semimembranosus (SM) and biceps femoris (BF), iliopsoas (IL), sartorius (SA) and rectus femoris (RF). For KFE, six electrodes are attached to six muscles: rectus femoris (RF), vastus medialis (VM) and vastus lateralis (VL), biceps femoris (BF), semimembranosus (SM) and semitendinosus (ST) [18]. The 6-axis force/torque sensors (SRI-M3203) are set at the robotic joints. The user's facial expression is detected in real time using the RealSense SR300-3D Camera. The experiment photograph is shown in Fig. 5.

During the experiments, the patients exert force on the FL to drive master robot, and the IL tries to actuate its own muscles to complete the gait task. The patients could relax their muscles between each trial. The reinforcement learning controller is trained and some hyperparameters are required, i.e. the learning rate is 0.001, the size of minibatch is 128, the discounting factor γ is 0.99, the communication time delay is $T_m = T_s = 50$ ms . Importantly, the weights for the each reward function term are modified with the relative entropy inverse reinforcement learning algorithm [14], and they are: the tracking error weight $\Lambda = \mathrm{diag}(10{,}10)$, the EMG weights for HFE: $\gamma_{GM} = 1.2$, $\gamma_{SM} = 0.9$, $\gamma_{BF} = 0.9$, $\gamma_{IL} = 1.2$, $\gamma_{SA} = 0.9$ and $\gamma_{RF} = 0.9$; the EMG weights for KFE: $\gamma_{RF} = 0.7$, $\gamma_{VM} = 0.7$, $\gamma_{VL} = 1.6$, $\gamma_{BF} = 0.7$, $\gamma_{SM} = 0.7$ and $\gamma_{ST} = 1.6$; the acceleration weight $\gamma_A = 0.5$, the action weight $\gamma_u = 0.3$. When the recognized facial expression is positive, V_{FER} =0.5; otherwise, when it is negative, V_{FER} =-3.5. The same reward function weights are shared across all five patients. After training the agent, the action is treated as the controller for the slave robot to conduct the experiments. The reward per episode and average reward in the training process are shown in Fig. 6. The reward increases quickly and converges at around 150 episodes and it takes about two hours for convergence when two robots work simultaneously, revealing high efficiency of the learning. Eventually, the average reward value reaches up to 290, and the learning stops. The training time can be further reduced by providing more rehabilitation robots for multiple patients' therapy at the same time for data collection, parallelizing the algorithm and pooling the policy updates asynchronously.

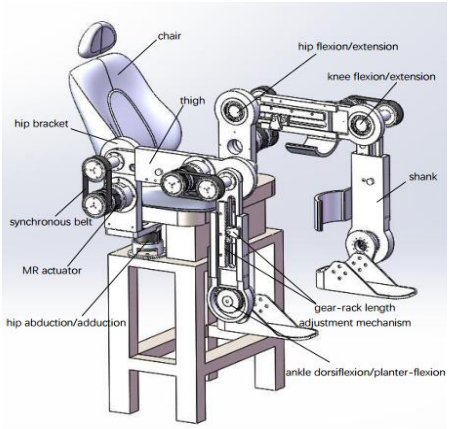


Figure 4. The robot design scheme.

TABLE I. Patients Information

Subject	Gender	Age	Years PS	FMA	Paretic Side
P1	male	56	2	24	left
P2	female	63	3	21	right
P3	male	59	3	20	left
P4	male	46	4	23	right
P5	female	66	8	16	left

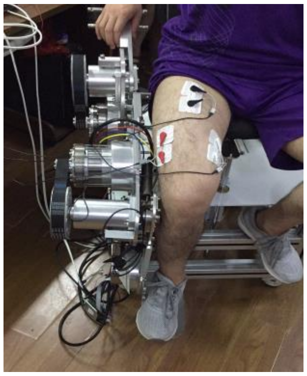


Figure 5. Experiment photograph.

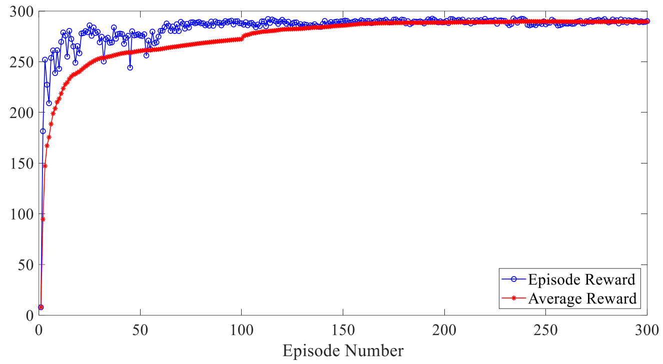


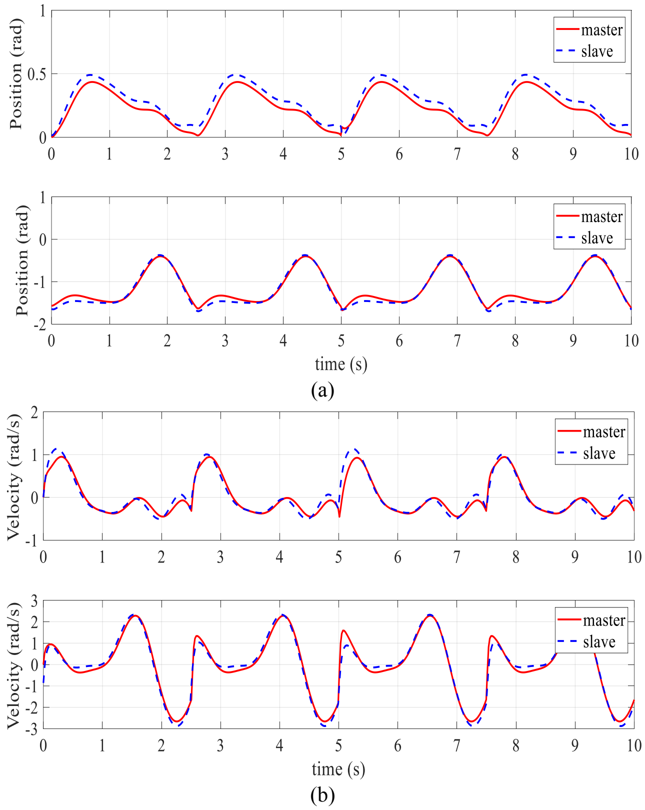
Figure 6. The episode and average reward versus the episode number during the reinforcement learning process.

In order to verify the control performance, the robotic joint positions and velocities are measured and presented in Fig. 7. It can be seen that the position tracking error and velocity tracking error are quite small throughout the experiments, validating the tracking effect of the learning controller. Also, some FER photos during the experiments are shown in Fig. 8, and the algorithm can differentiate clearly.

Clinical tests are carried out with the proposed framework, where the five hemiplegic patients (P1-P5) are scheduled to conduct 90-minute therapy four times a week for eight weeks. During each training, the physiotherapists would suggest the exercise matching their rehabilitation situation and supervise the robot-assisted therapy. To increase interest, some VR games, like reaching task and kicking football, are involved. FER proves necessary in the tests, especially when the patients show painful facial expression and the robot can stop immediately to avoid further injury. By contrast with the same control framework without FER, the movement tracking difference is not obvious, but the normalized relative EMG values get higher by around 10% and the patients rehabilitate sooner. After completing the whole therapy, the FMA is employed to evaluate the rehabilitation performance, and the scores are recorded by multiple trials. The five patients' FMA scores turn to 33, 31, 30, 29 and 27 respectively, and the average increase is 9.2, which is high enough in 2 months and proves that the proposed reinforcement learning-based mirror therapy can efficiently improve the rehabilitation efficacy.

Afterwards, we apply the bilateral impedance control [4] to do the same test with another group of five hemiplegic patients, and the comparison between the learning control and the impedance control without learning is listed in Table II. It can be found that the mean position tracking error (P=0.03) and mean velocity tracking error (P=0.04) with learning are both less than those without learning (28.57% for HFE position tracking error, 33.3% for KFE position tracking error, 43.1% for HFE velocity tracking error, 42.86% for KFE velocity tracking error). This indicates that the position and velocity tracking effect for the learning control is better than the one without learning. The reason is that disturbance and uncertainties exist in the robot dynamic model and impedance model, while they can be eliminated with the help of neural networks in the deep reinforcement learning control. Besides, the normalized relative EMG values (E_{ILi}/E_{FLi}) of the main muscles (GM, IL, VL and ST) under the two control conditions are recorded. The differences in the muscle activity for the four muscles are all significant (P=0.006 for GM, P=0.008 for IL, P=0.012 for VL, P=0.013 for ST), and the improvement of the mean normalized relative EMG values are great (0.19 for GM, 0.34 for IL, 0.23 for VL, 0.37 for ST). In other words, the patients under the learning control own much higher muscle activation than the one without learning. The reinforcement learning control is able to adjust the exercise intensity instantly according to the subject's rehabilitation situation, and it can suit for various subjects with different exercise abilities after learning from demonstrations. While the impedance model parameters are difficult to determine and vary largely among different subjects, it takes much time for the controller to adapt to the patients, and the training modality is monotonous. So, within the certain therapy period, the control strategy with learning can obtain better rehabilitation efficacy than the impedance control. Moreover, the mean increase of FMA score for five patients after therapy is taken down in Table II. The increase of FMA Score for the method with learning is 9.2, while it is only 6.2 for the approach without learning, which further certifies that the learning control outperforms the impedance control without learning.

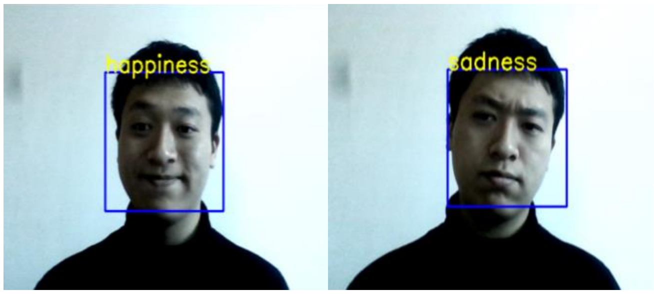
Compared with other state-of-the-art control strategies, the proposed reinforcement learning control aims to maximize the IL's muscle activation and ensure that its motion can be mastered by the FL at the same time. The subject's emotion is also incorporated in the framework. These all can be realized cooperatively by setting the proper reward function. Besides, the uncertainties in the robot dynamics can be eliminated due to its model-free characteristic. Another advantage is that the proposed approach has the potential to deal with different kinds of patients with different levels of movement disabilities.



(a)

(b)

Figure 7. The tracking effect between the master and slave robot during the experiment procedure, where the above figure denotes HFE and the below figure denotes KFE. (a) The position of the master and slave robot. (b) The velocity of the master and slave robot.



(a) (b)

Figure 8. The facial expression recognition photos during the experiment procedure. The recognized results are (a) happiness and (b) sadness.

TABLE II. CONTROL PERFORMANCE COMPARISON

Metrics	***Without Learning***	***With Learning***
Position Tracking Effect		
Mean Error for HFE	0.07 rad	0.05 rad
Mean Error for KFE	0.12 rad	0.08 rad
Velocity Tracking Effect		
Mean Error for HFE	0.58 rad/s	0.33 rad/s
Mean Error for KFE	1.05 rad/s	0.60 rad/s
Normalized Relative EMG Values		
GM	0.33±0.07	0.52±0.03
IL	0.34±0.05	0.68±0.09
VL	0.42±0.07	0.65±0.11
ST	0.44±0.08	0.81±0.15
Rehabilitation Efficacy		
Increase of FMA Score	6.2	9.2

VII. CONCLUSION

This paper has presented a bilateral master-slave robotic system for hemiparesis rehabilitation. The master robot is interacted with the FL, which is controlled with the impedance model. The slave robot is wearable for the IL, and its controller is based on reinforcement learning. Signals through multiple channels, including the motion trajectory, muscle activation and the user's emotion, are combined in the learning algorithm. The learning aims to maximize the rehabilitation efficacy and minimize the trajectory tracking error through mirror transmission. Also, the user's positive facial expression is encouraged in the scheme. A lower extremity rehabilitation robot with MR actuators has been designed and implemented. The experiments using the robot are completed, and the rehabilitation performance is satisfactory. In the future work, the classifications of the facial expression will be reduced and detection of pain level through FER will be emphasized to enhance the reliability and robustness. More types of biological information, such as electrocardiogram, skin conductance response and respiration, will be fused in the reinforcement learning framework. Moreover, functional electrical stimulation will be included to further improve the muscle contraction ability.

REFERENCES

[1] Mahya Shahbazi, Seyed Farokh Atashzar, et al., "Robotics-assisted mirror rehabilitation therapy: a therapist-in-theloop assist-as-needed architecture," *IEEE/ASME Transactions on Mechatronics*, vol. 21, no. 4, pp. 1954-1965, August 2016.

[2] Mojtaba Sharifi, Saeed Behzadipour, et al., "Cooperative modalities in robotic tele-rehabilitation using nonlinear bilateral impedance control," *Control Engineering Practice*, vol. 67, pp. 52-63, 2017.

[3] Masashi Hamaya, Takamitsu Matsubara, et al., "Learning assistive strategies from a few user-robot interactions: model-based reinforcement learning approach," in *Proceedings of IEEE International Conference on Robotics and Automation*, Stockholm, Sweden, May, 16-21, 2016, pp. 3346-3351.

[4] Hamidreza Modares, Isura Ranatunga, Frank L. Lewis, et al., "Optimized assistive human-robot interaction using reinforcement learning," *IEEE Transactions on Cybernetics*, vol. 46, no. 3, pp. 655-667, March, 2016.

[5] Shixiang Gu, Timothy Lillicrap, et al., "Continuous deep Q-learning with model-based acceleration," arXiv:1603.00748.

[6] Panagiotis Paraskevas Filntisis, Niki Efthymiou, et al., "Fusing body posture with facial expressions for joint recognition of affect in child-robot interaction," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp.4011-4018, October, 2019.

[7] Mojtaba Sharifi, Hassan Salarieh, et al., "Impedance control of nonlinear multi-dof teleoperation systems with time delay: absolute stability," *IET Control Theory and Application*, 2018.

[8] Ulysse Côté-Allard, Cheikh Latyr Fall, et al., "Deep learning for electromyographic hand gesture signal classification using transfer learning," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 4, pp. 760-771, April, 2019.

[9] Joseph M. Hidler, Marti Carroll and Elissa H. Federovich, "Strength and coordination in the paretic leg of individuals following acute stroke," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 15, no.4, pp. 526-534, 2007.

[10] Wei Sin Ang, Hartmut Geyer, et al., "Objective assessment of spasticity with a method based on human upper limb model," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 7, pp. 1414-1423, 2018.

[11] P. Lucey, J. F. Cohn, et al., "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Proceedings of IEEE Computer Vision and Pattern Recognition Workshops*, 2010, pp. 94–101.

[12] Behzad Hasani and Mohammad H. Mahoor, "Facial expression recognition using enhanced deep 3D convolutional neural networks," in *Proceedings of IEEE Computer Vision and Pattern Recognition Workshops*, 2017.

[13] Christian Szegedy, Sergey Ioffe and Vincent Vanhoucke, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," in *Proceedings of IEEE Computer Vision and Pattern Recognition*, 2016.

[14] Abdeslam Boularias, Jens Kober and Jan Peters, "Relative entropy inverse reinforcement learning," in *Proceedings of Artificial Intelligences and Statistics*, 2011, pp. 20-27.

[15] Jiajun Xu, Youfu Li, et al., "A multi-mode rehabilitation robot with magnetorheological actuators based on human motion intention estimation," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 10, pp. 2216-2228, 2019.

[16] Jiajun Xu, Linsen Xu, et al., "Design and implementation of the lower extremity robotic exoskeleton with magnetorheological actuators," in *Proceedings of IEEE International Conference on Mechatronics and Automation*, Tianjin, China, August 4-7, 2019, pp. 1294-1299.

[17] Jiajun Xu, Linsen Xu, et al., "Design of lower extremity rehabilitation robots with magnetorheological dampers and wire-driven system," in *Proceedings of IEEE International Conference on Information and Automation*, Wuyi Moutain, China, 2018, pp. 395-400.

[18] L. M. Schutte, M. M. Rodgers, et al., "Improving the efficacy of electrical stimulation-induced leg cycle ergometry: An analysis based on a dynamic musculoskeletal model," *IEEE Transactions on Rehabilitation Engineering*, vol. 1, no. 2, pp. 109–125, July, 1993.