Natural teaching for humanoid robot via human-in-the-loop scene-motion cross-modal perception

Wenbin Xu
School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai, China
Xudong Li, Liang Gong and Yixiang Huang
Shanghai Jiao Tong University, Shanghai, China
Zeyuan Zheng
School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai, China
Zelin Zhao
Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
Lujie Zhao, Binhao Chen and Haozhe Yang
Shanghai Jiao Tong University, Shanghai, China
Li Cao
Shanghai Jiao Tong University Medical School Affiliated Ruijin Hospital, Shanghai, China, and
Chengliang Liu
Shanghai Jiao Tong University, Shanghai, China

Abstract

Purpose – This paper aims to present a human-in-the-loop natural teaching paradigm based on scene-motion cross-modal perception, which facilitates the manipulation intelligence and robot teleoperation.

Design/methodology/approach – The proposed natural teaching paradigm is used to telemanipulate a life-size humanoid robot in response to a complicated working scenario. First, a vision sensor is used to project mission scenes onto virtual reality glasses for human-in-the-loop reactions. Second, motion capture system is established to retarget eye-body synergic movements to a skeletal model. Third, real-time data transfer is realized through publish-subscribe messaging mechanism in robot operating system. Next, joint angles are computed through a fast mapping algorithm and sent to a slave controller through a serial port. Finally, visualization terminals render it convenient to make comparisons between two motion systems.

Findings – Experimentation in various industrial mission scenes, such as approaching flanges, shows the numerous advantages brought by natural teaching, including being real-time, high accuracy, repeatability and dexterity.

Originality/value – The proposed paradigm realizes the natural cross-modal combination of perception information and enhances the working capacity and flexibility of industrial robots, paving a new way for effective robot teaching and autonomous learning.

Keywords Humanoid robot, Cross-modal perception, Human-in-the-loop, Motion imitation, Natural teaching

Paper type Research paper

1. Introduction

In recent years, demands for industrial robots with high intelligence have shown a tremendous growth in military, medicine, manufacturing and social life. Industrial robots are increasingly faced up with challenges of executing complicated tasks in unstructured environments, such as welding tracking on a curved surface, sorting and placing of scattered workpieces with surfaces of multiple types, like

The current issue and full text archive of this journal is available on Emerald Insight at: **www.emeraldinsight.com/0143-991X.htm**

Industrial Robot: the international journal of robotics research and application
46/3 (2019) 404–414
Emerald Publishing Limited [ISSN 0143-991X]
[DOI 10.1108/IR-06-2018-0118]

This study was supported by the Natural Scientific Foundation of China under Grant No. 51775333 and the Shanghai Jiao Tong University Foundation of Cross-disciplinary Research on Medicine and Engineering under Grant No. YG2016MS64.

Received 9 June 2018
Revised 24 August 2018
5 October 2018
8 November 2018
16 November 2018
Accepted 12 January 2019

three-way valves and flanges and so on. Many paradigms are adopted to improve the ability of robots to perform complex tasks based on data-driven methods (Deisenroth *et al.*, 2013; Bohg *et al.*, 2014; Mnih *et al.*, 2015; Levine *et al.*, 2018). However, with limited data, those data-driven methods alone tend to have a poor performance. Under such a circumstance, the combination of teaching and machine learning to cope with the lack of data has achieved good results (Koenig and Matarić, 2017; Finn *et al.*, 2017; Michini *et al.*, 2013; Kassahun *et al.*, 2016; Osentoski *et al.*, 2012). As a direct way to endow industrial robots with human's knowledge, teaching renders the intelligence development of robots more than possible.

As a matter of fact, traditional teaching methods are faced with multiple difficulties. First, when teaching is performed for complicated motions with multiple degree of freedom (DOF), an expert is necessary for demonstration and the effect of teaching highly depends on his knowledge. The numerous frames of continuous movements will cause a sharp increase in the amount of teaching information, thus placing a heavy burden on the expert (Silver *et al.*, 2012; Laskey *et al.*, 2016; Koert *et al.*, 2016). Second, to facilitate robots to understand human's teaching information and to develop intelligence, behavior recognition and semantic classification are necessary, while conventional demonstration methods often neglect the transmission of semantic information (Wächter and Asfour, 2015; Lim, 2016; Rozo *et al.*, 2016; Alibeigi *et al.*, 2017). Third, the ability to make decisions based on multiple kinds of sensory information is an important manifestation of human intelligence while conventional demonstration methods generally overlook or misunderstand the relations between different sensory information (Lim and Okuno, 2014; Noda *et al.*, 2014).

Considering children's learning process, they observe the behavior of adults and then reproduce it (Riley *et al.*, 2003). Such a process is always natural and effective because human beings share the same way to comprehend scenes and the same behavioral language. Inspired by this, natural teaching is the key to overcoming obstacles to the exchange of teaching information between human and robots. Natural teaching is actually a branch of human-robot interaction (HRI) technology, representing a kind of teaching paradigm which is user-friendly and coordinates human and robot in scene comprehension. Aimed at completing tasks with specified human semantic information, natural teaching is an end-to-end and highly efficient method for interaction with surroundings or complicated movements. Moreover, fulfilling such tasks is conducive to establishing a deep understanding of potential implications from training data through subsequent intelligence algorithms, thus achieving a high level of intellectual development.

Scene-motion cross-modal perception constitutes a critical component of natural teaching. Enlightened by role-playing in E-sports, the demonstrator is provided with visual information to perceive the mission scenario of the robot and then implements movements. The demonstrator's eye-body synergic movements are collected as motion information. Thanks to VR (Virtual Reality) and HRI technology, the robot and the demonstrator can share the common visual and motion information during the whole teaching process. Through such a perception system, the robot can achieve the cross-modal combination of scene and motion information, while the demonstrator can have an overall cognition of the surroundings from first person view (FPV), analyze complicated information and make movement decisions. Furthermore, the recording of intricate multi-DOF movements, as well as the live video stream, provides the robot with comprehensive scene-motion information so that the robot can be gradually endowed with the ability to repeat the same process. In the near future, the robot can even develop the capability of making autonomous decisions through such a natural teaching paradigm.

Employing humanoid robots as a platform to verify the natural teaching paradigm with scene-motion cross-modal perception can provide numerous advantages. First, since humanoid robots possess human-like structures and scales that have evolved for millions of years, the abundant DOF and the complex joints between links of a humanoid can represent an industrial robot with an extremely complicated structure. Second, the excellent mobility potential of humanoid robots renders it possible for them to be assigned with different tasks (Koenemann *et al.*, 2014). Third, humanoids can serve as a direct and natural platform for natural teaching. As they can completely reflect human motion, demonstrators can easily assess the difference between human motion and robot imitation during motion synchronization. Demonstrators can further consider the conversion of postures from human to the robot and optimize the conversion rules against corresponding problems (Argall *et al.*, 2009).

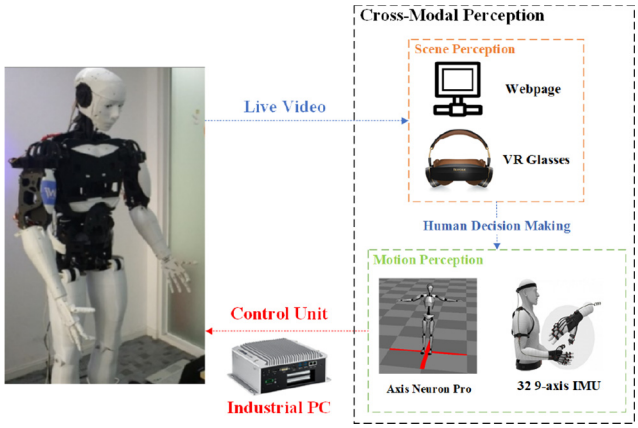
Herein we report a human-in-the-loop natural teaching paradigm with motion-scene cross-modal perception on a life-size humanoid robot. The robot is established based on InMoov, an open-sourced 3D printing humanoid robot. It possesses 29 DOF, 22 of which are controlled in this system. The following is the natural teaching process. First, a vision sensor is employed to project the mission scene onto the VR glasses. Second, motion perception captures the motion of human with a set of wearable sensors and presents the collected motion data in BVH (BioVision Hierarchy) format. The motion data are transmitted to an industrial PC (IPC) through TCP/IP and parsed according to BVH format. Next, the parsed Euler angles are converted to corresponding joint angles through a fast mapping algorithm and encapsulated in a communication protocol. At last, IPC sends joint angles to the slave controller to control the robot. The whole system has paved a novel, real-time and accurate way for a natural teaching paradigm on humanoid robots.

This paper is organized as follows. In Section 2, the scene-motion cross-modal perception system is introduced. Section 3 discusses the setup of the humanoid robot. Section 4 presents the scheme of real-time motion imitation on the humanoid. Section 5 performs several experiments based on the proposed natural teaching paradigm. Finally, section 6 deals with the conclusion about our work.

2. Scene-motion cross-modal perception

The framework of the cross-modal perception system is shown in Figure 1. Scene perception makes it possible for the demonstrator to remotely perceive the complicated surroundings around the robot, while motion perception passes back real-time human motions to the controller. The combination of scene and motion perceptions takes full advantage of human's intelligence because

Figure 1 Scene-motion cross-modal perception system



each movement in the loop is determined by human and reflected on the robot.

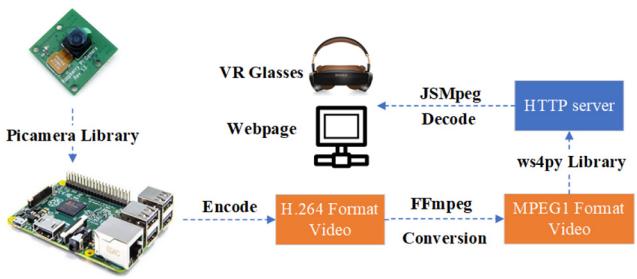
2.1 Scene perception

Scene perception is achieved through a remote video stream and multiple display terminals. Figure 2 shows its principle and the video stream. The main function of scene perception is to stream the live video recorded by Raspberry Pi Camera to the demonstrator. Since the camera is installed in one eye of the robot, the demonstrator wearing VR glasses can make decisions about movements from FPV. Besides, multiple display terminals make it possible for users to watch the same video stream on different electrical devices.

2.1.1 Remote video stream

Raspberry Pi is selected as the processing unit to drive the Pi camera for remote live video monitoring. The video obtained from Raspberry Pi is encoded in H.264 format, which is barely supported in browsers. Hence, FFmpeg (Fast Forward mpeg) is adopted to convert the H.264 format to the MPEG1 format. The video stream is then uploaded to an HTTP server through ws4py. Decoding is completed with JSMpeg, an excellent MPEG1 video and MP2 audio decoder defined in JavaScript. At most 30 fps video with a resolution of 1280×960 can be decoded by JSMpeg. Since JSMpeg is based on JavaScript, the video stream works in most modern browser (i.e. Firefox, Edge, Chrome, etc). Moreover, the decoder has a low latency via WebSockets, thus achieving the real-time feature of our work.

Figure 2 Principle of scene perception



2.1.2 Multiple display terminals

Display terminals include VR glasses and webpages. The type of VR glasses we adopt is Royole Moon, a combination of a headset, vari-focusing glasses, and a control terminal. The operating system of Royole Moon is Moon OS developed based on Android. What's more, it provides free access to the external network, which means users can access the live video stream and perceive the mission scene at a distance. However, since the video capture is accomplished using one camera, all videos are two-dimensional. It is inevitable that some necessary information will be lost if the demonstrator watches the screen alone. Therefore, some external assistance is required to improve the user experience. For webpage terminals, the principle is basically the same with VR glasses. Any devices which have installed a modern browser are accessible to the low-latency live video stream through a specified URL.

2.2 Motion perception

To capture motion information, several methods have been adopted. Gobee *et al.* (2017) fixes IR sensors and accelerometer motion sensors to human legs and achieves real-time control of gaits on a biped humanoid robot. Durdu *et al.* (2014) attaches potentiometers to human joints and then collect motion data. Furthermore, vision sensing technology is also employed. Several articles (Yavşan and Uçar, 2016; Ding *et al.*, 2014; Bindal *et al.*, 2015) utilize Kinect for gesture recognition and then perform similar actions on robots through diverse algorithms. Herein motion recording is achieved through wearable sensors. The motion capture system is composed of a motion sensor to capture real-time human motion and a human motion retargeting method.

2.2.1 Motion sensor

A modular system composed of 32 9-axis sensors is adopted as the motion sensor. It is a set of wearable sensors designed by Noitom Technology Ltd. to deliver motion capture technology. It contains 32 IMUs (Inertial Measurement Unit), each of which is composed of a 3-axis gyroscope, 3-axis accelerometer, and 3-axis magnetometer. The static accuracy of each IMU is ± 1 degree for roll/pitch angle and ± 2 degree for yaw angle. The system is operated with Axis Neuron Pro (ANP) running on Windows OS for calibration and management. In addition, a skeleton model is visualized in ANP to reflect real-time human motion. Another important feature of ANP is the function to broadcast BVH data through TCP so that other programs can obtain and analyze these data.

2.2.2 Human motion retargeting

Motion retargeting is a classic problem which aims to retarget motion from one character to another while keeping styles of the original motion (Meng *et al.*, 2017). With this method, real-time human motion can be displayed on the skeletal model in ANP through BVH data. As a universal file format for human feature animation usually adopted in skeleton models, it can store motion for a hierarchical skeleton, which means that motion of the child node is directly dependent on the motion of the parent one (Dai *et al.*, 2010). As shown in Figure 3, a normal BVH file will consist of several parts as follows.

- HIERARCHY signifies the beginning of skeleton definition.
- ROOT defines the root of the whole skeleton.

Figure 3 An Example of BVH Format

```
HIERARCHY
ROOT Hips
{
    OFFSET 0.00 104.19 0.00
    CHANNELS 6 Xposition Yposition Zposition Yrotation Xrotation Zrotation
    JOINT RightUpLeg
    {
        OFFSET -11.50 0.00 0.00
        CHANNELS 6 Xposition Yposition Zposition Yrotation Xrotation Zrotation
        JOINT RightLeg
        {
            OFFSET 0.00 -48.00 0.00
            CHANNELS 6 Xposition Yposition Zposition Yrotation Xrotation Zrotation
            JOINT RightFoot
            {
                OFFSET 0.00 -48.00 0.00
                CHANNELS 6 Xposition Yposition Zposition Yrotation Xrotation Zrotation
                End Site
                {
                    OFFSET 0.00 -1.81 18.06
                }
            }
        }
    }
MOTION
Frames: 2
Frame Time: 0.04166667
-9.533684 4.447926 -0.566564 -7.757381 -1.735414 89.207932 9.763572
```

- OFFSET specifies the deviation of the child joint from its parent joint, which remains constant.
- CHANNELS contains several parameters. The first parameter indicates the number of DOF. Usually, only the root joint has both position data and rotation data. The rest ones only contain rotation data in the form of Euler angles. The sequence of rotation hinges on the sequence mentioned in CHANNELS, i.e. the rotation is carried out in YXZ order in the example.
- End Site is only tagged in the definition of an end-effector and describes the lengths of bones through OFFSET.
- MOTION represents the beginning of another section which describes the states of each joint at each frame.
- Frame Time is the duration of each frame. The rest data are real-time states of each joint described sequentially in the HIERARCHY section. Hence, the number of these data is equal to the that of channels defined in the HIERARCHY section.

We adopt BVH with no position channels since position values keep constant. Hence, three rotation values are obtained for each joint. Accordingly, we can describe human postures through these Euler angles based on the assumption that wearable sensors keep fixed with respect to human body.

3. Setup of the humanoid robot

To realize real-time motion imitation, a humanoid robot is assembled since it possesses human-like design and is able to mimic human motion (Rodriguez *et al.*, 2006). However, due to the complicated structure of the robot and various constraints of conventional manufacturing methods, it is difficult to fulfill an elegant design of a dexterous humanoid robot. Fortunately, with the rapid advancement in additive manufacturing, 3D printing turns to be more cost-effective. Moreover, 3D printing element is also becoming stronger, more accurate and therefore more reliable. 3D-printed humanoid robots like InMoov, Flobi and iCub have been created to serve as experiment platforms where research on HRI is conducted.

Here a 3D-printed life-size humanoid robot is established based on InMoov initiated by Langevin (2014), a French sculptor in 2012. The whole structure, as well as other necessary backgrounds, have been illustrated in the previous work (Gong *et al.*, 2017). 22 out of 29 DOF are controlled during motion imitation, including 5 DOF for each hand, 2 for each arm, 3 for each shoulder and 2 for the neck, as shown in Figure 4. As for control, the slave controller is composed of 4 small Arduino Nano control core boards, each of which can drive 6 servos with corresponding angles through PWM, and an Arduino Mega 2560 master board which communicates with the aforementioned nano nodes via 485 Hub based on the Modbus RTU control.

4. Real-time imitation of human motion

The whole structure of the proposed method is shown in Figure 5. First, the publish-subscribe messaging mechanism and the designed communication protocol ensures the security of data transfer. Second, the fast mapping algorithm converts BVH data into corresponding joint angles. Next, visualization terminals enable users to make comparisons between different but simultaneous motion systems.

4.1 Data transmission

During data transmission, communication protocols and quantization are necessary to prevent undesirable communication delays and packet dropouts (Liu *et al.*, 2016). Herein the publish-subscribe messaging mechanism and a specific protocol are designed to realize the reliable data transmission. To be more

Figure 4 DOF of humanoid robot (DOF of fingers are not displayed)

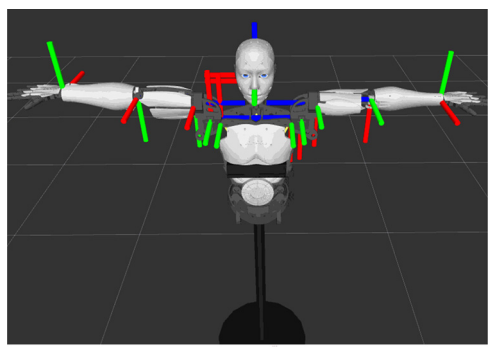
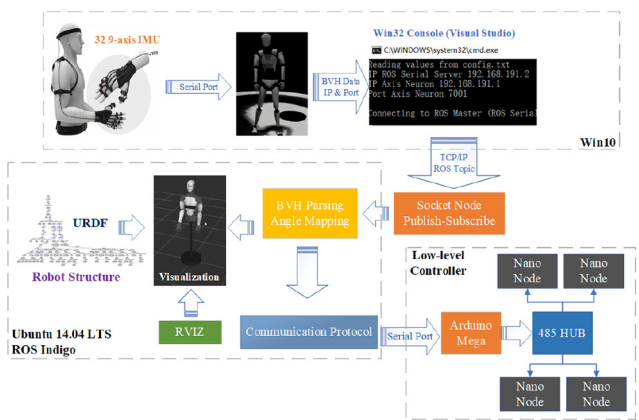


Figure 5 Whole structure of proposed method



specific, the publish-subscribe messaging mechanism allows nodes, which are executables after compilation, to publish messages or subscribe to a topic (Wang *et al.*, 2016). Topics are asynchronous and highly efficient. The whole data stream is mainly enabled through such a messaging mechanism, as shown in Figure 6, where ellipses stand for nodes and squares represent topics:

- Socket_node connects with the win32 console through TCP/IP and then advertises the topic, PN_node/data.
- Mapping_node subscribes to the previous topic and then converts BVH data to joint angles, which are then published to another topic called Joint_angle.
- joint_state_publisher achieves the real-time simulation of the robot model using the calculated joint angles.
- Serial_node realizes the serial communication between the master and slave computers.

While topics have been successfully implemented in the data transfer process, reliable communication between the master and slave computers is still necessary to control the robot. Before transmission, all these data including a time stamp and joint angles are quantized to integers. The communication protocol contains 2 bits of time stamp data, 22 bits of position data corresponding to each joint, and 2 bits of CRC16 check code which are generated according to prior 27 bits to ensure the safety of data transfer, as shown in Figure 7.

4.2 Mapping algorithm

Several methods have been adopted to achieve motion imitation. Riley *et al.* (2003) computes joint angles through a fast full-body inverse kinematics (IK) method. The full-body IK problem is divided into many sub-problems to realize real-time imitation on a Sarcos humanoid robot with 30 DOF. Koenemann *et al.* (2014) realizes complex whole-body motion imitation on a Nao humanoid based on the positions of end-effectors and center of mass. By actively balancing the center of mass over the support polygon, the proposed approach enables the robot to stand on one foot as the demonstrator does. Durdu *et al.* (2014) classifies the collected data with the assistance of ANN to perform movements on the robot. Herein, a fasting mapping algorithm is employed to realize the transformation.

To make the robot imitate human motion, the key point is how to compute the corresponding joint angles from BVH data. BVH has provided us with three euler angles for each node, enabling us to ascertain the rotation matrix between child and parent links. Denote euler angles with a rotation order of ZYX as φ, θ, ψ, the rotation matrix of child frame with respect to parent frame is:

$$R_{child}^{parent} = \begin{pmatrix} \cos\varphi & -\sin\varphi & 0 \\ \sin\varphi & \cos\varphi & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\psi & -\sin\psi \\ 0 & \sin\psi & \cos\psi \end{pmatrix} \tag{1}$$

To describe the motion quantitatively, here we consider human motion as a sequence of rotation matrices. f_i is the rotation matrix at BVH frame i:

$$f_i = \left\{ R_{LHand}^{LForearm}, R_{LForearm}^{LArm}, R_{LArm}^{Body}, R_{Head}^{Body}, R_{RHand}^{RForearm}, R_{RForearm}^{RArm}, R_{RArm}^{Body} \right\} \tag{2}$$

Thus, each posture is defined as a sequence of rotation matrices at frame i, i.e. $R_{LHand}^{LForearm}$ stands for the rotation matrix between left hand and left forearm. Similarly, we can also define robot motion as another sequence. The goal is to eliminate the difference between each corresponding rotation matrix of human and robot as much as possible. Figure 8 states the mapping problem. On one hand, human, with physiological constraints, cannot have 3 rotational DOF at each joint and some of them are not completely independent. On the other hand, due to mechanical constraints, many joints of the humanoid are also unable to rotate in three independent directions. Hence, each joint requires respective discussion for the mapping algorithm. Thanks to the structural symmetry, the algorithms for $R_{Ljoint1}^{Ljoint2}$ and $R_{Rjoint1}^{Rjoint2}$ share the same principle.

4.2.1 Shoulder joint

The first case is the mapping between shoulders, which entails conversion from 3 human DOF to 3 robot DOF. Three rotational joints are installed on each shoulder part of InMoov and their axes of rotation can be approximately treated as perpendicular to each other. Denote the joint angles of 3 shoulder joints as respectively α, β, γ and the rotation matrix of the arm link with respect to the body can be similarly expressed as:

$$R_{Arm}^{Body} = \begin{pmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cos\beta & 0 & \sin\beta \\ 0 & 1 & 0 \\ -\sin\beta & 0 & \cos\beta \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\gamma & -\sin\gamma \\ 0 & \sin\gamma & \cos\gamma \end{pmatrix} \tag{3}$$

With equations (1) and (3) , we can derive a one-to-one correlation:

$$\alpha = \varphi, \beta = \theta, \gamma = \psi \tag{4}$$

Figure 6 Visualized data stream

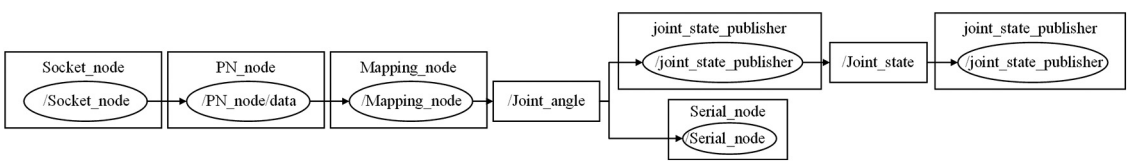
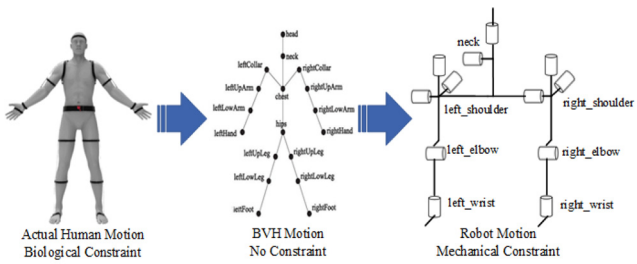


Figure 7 Designed communication protocol

Figure 8 Three motion systems with different constraints



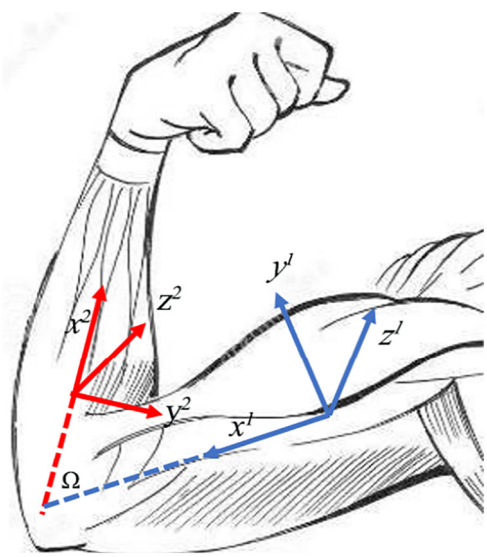
But there's still one thing that needs to be noticed. Since the mechanical structure of the robot has determined the rotation order of three joints between body and arm, the rotation order of euler angles in BVH should be set to be the same, which in our case is ZYX.

4.2.2 Elbow joint

The mapping between elbow joints entails the conversion from 2 human DOF to 1 robot DOF. Human elbows are able to bend and rotate while those of the robot can only bend. Then we need to compute the joint angle for bending, which is Ω, as shown in Figure 9. With the assumption that sensors are fixed with respect to the human body and the x-direction is along the forearm link, we can derive the following equations:

$$\hat{x_2}^2 = (1, 0, 0)^T \tag{5}$$

Figure 9 Elbow mapping



$$\hat{x_2}^1 = R_2^1 \hat{x_2}^2 = (\cos\varphi\cos\theta, \cos\varphi\sin\theta, -\sin\theta)^T \tag{6}$$

$$<\hat{x_2}^1, \hat{x_1}^1> = \arccos(\cos\varphi\cos\theta) \tag{7}$$

$$\Omega = \pi - <\hat{x_2}^1, \hat{x_1}^1> \; = \pi - \arccos(\cos\varphi\cos\theta) \tag{8}$$

R_2^1 stands for the rotation matrix of frame $x_2y_2z_2$ with respect to $x_1y_1z_1$. $\hat{x_1}^1$ is a unit vector of x_1 in frame $x_1y_1z_1$.

4.2.3 Neck joint

Mapping between neck joints requires the conversion from 3 human DOF (φ, θ, ψ) to 2 robot DOF (α, β). Due to the mechanical constraints, only rotations in two directions can be retained. The solution to this case resembles that for the shoulder joint and should be written as

$$\alpha = \varphi, \beta = \theta \tag{9}$$

4.3 Visualization terminals

On one hand, in the display of human motion collected from the motion sensor, ANP can visualize the aforementioned skeletal model, where each joint possesses three DOF despite human's physiological constraints.

On the other hand, to visualize motion on the humanoid and to make the simulation more convenient, another visualization scheme is provided with the assistance of robot operating system (ROS). A 3D visualization model is created in URDF (Unified Robot Description Format), a language based on XML and designed to describe the robot simulation model universally in ROS system, including the shape, size, and color, kinematic and dynamic characteristics of the model. Wang *et al.* (2016) has introduced the basic grammars. However, the highly repetitive mechanical structure of InMoov makes it arduous to write a URDF manually. Hence we resort to a powerful tool called Xacro (XML Macros). Xacro is adopted to reuse the same structure for two different parts, i.e. left arms and right arms and to auto-generate a URDF file. Fundamental grammars are shown in Table I. After importing STL files with scale adjustments, the robot model can operate with the computed joint angles in RVIZ, a 3D visualization tool in ROS. These two visualization terminals are shown in Figure 10.

5. Experiment

This section designs several experiments to demonstrate the accuracy, repeatability, and dexterity of the proposed natural teaching paradigm and discusses the experimental results.

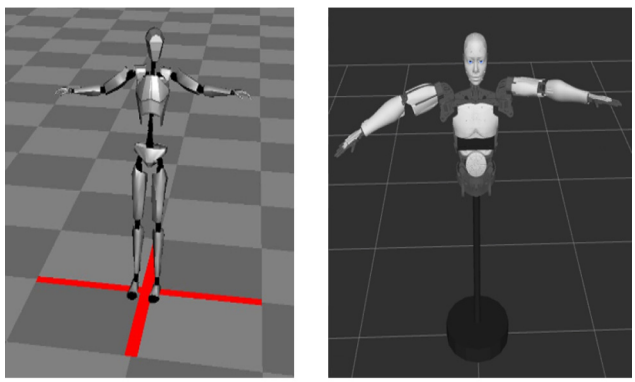
5.1 Accuracy

First, the accuracy of the control method is verified with several motion imitation experiments. Snapshots of postures are taken, including various positions of two arms, face orientations and movements of fingers. The results can be examined in Figure 11 and Figure 12. For these complicated gestures, the high degree of similarity between the demonstrator and the humanoid robot has proven that the robot can successfully follow the upper limb motion of the demonstrator, thus reflecting the feasibility and accuracy of our proposed method. Moreover, the synchronous

Table I Fundamental grammars of Xacro

Command	Definition	Usage
Property	<xacro:property name="pi" value="3.14"/>	< · · · value = "${2*pi}" · · ·/ >
Argument	<xacro:arg name="use_gui" default="false"/ >	< · · ·\ use_gui:= true · · ·/ >
Macro	<xacro:macro name="arm" params="side"/ >	<xacro:arm side="left"/ >
Including	<xacro:include filename="other_file.xacro"/>	

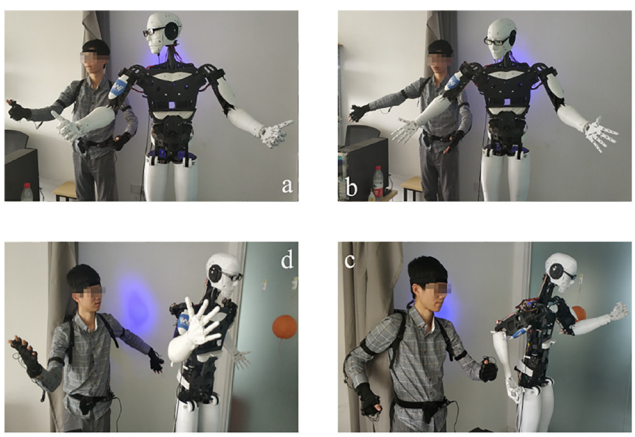
Figure 10 Different visualization terminals for different motion systems



(a) **(b)**

Notes: (a) Skeletal Model in ANP; (b) robot Model in RVIZ

Figure 11 Experiments of different gestures



latency of fewer than 0.5 seconds validates the real-time performance.

To further illustrate accuracy, the second experiment is carried out to measure the angle errors for individual motions. Figure 13 shows the rotation directions and initial gestures. The angle errors for three motions are measured and plotted in Figure 14 and it follows $Error = |Angle_{Robot} - Angle_{Human}|$. The generally small errors are acceptable for natural teaching and further prove that the motion imitation system possess high accuracy.

Figure 12 Comparison between Fingers

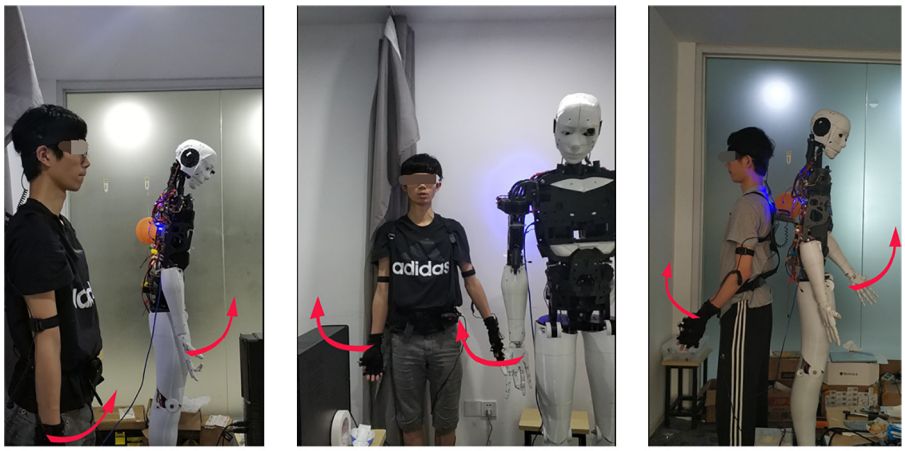
However, there are still some limitations. The first one is the difference between the structures of human and robot. Each of our arms has 7 DOF but the robot has only 5 and the rotational axes of the wrists are not the same. Besides, for some joints, the range of movement is limited due to its mechanical constraints. The second one is the mismatch between the skeletal model visualized through BVH data and the demonstrator's real motion. Revolution of each joint is achieved through skeletons in the human body, while the wearable sensors can only remain fixed to the skin or clothes. The angular displacements between our skin and skeletons cause the measurement error. Other factors include the accumulated drift errors and different positions of wearable sensors relative to human bodies. Nevertheless, there are still some possible solutions to these limitations. For example, sensors can be bound tightly to limbs in case of relative displacement. Human motion can be confined to a certain range to achieve a higher accuracy. Reasonable compensations for angular error can also be designed to render the motion retargeting more reliable.

5.2 Repeatability

After the demonstration is completed, the robot is expected to repeat the same motion, which means high repeatability is desired. To verify the repeatability of the proposed natural teaching paradigm, an experiment is carried out where the robot is expected to approach the same point with its index finger. The distance errors are listed in Table II. The average distance error is 6.8mm, which is relatively small compared to its size. Aging of actuators, frictions in transmission mechanisms and instability of power supply may contribute to these gross distance errors.

To further illustrate the repeatability, another more complicated teaching experiment is carried out. Industrial robots are often required to repeat precise operations. Hence,

Figure 13 Snapshots for three individual motions



Notes: (a) Shoulder Up; (b) shoulder side; (c) elbow

Figure 14 Angle error between human and robot

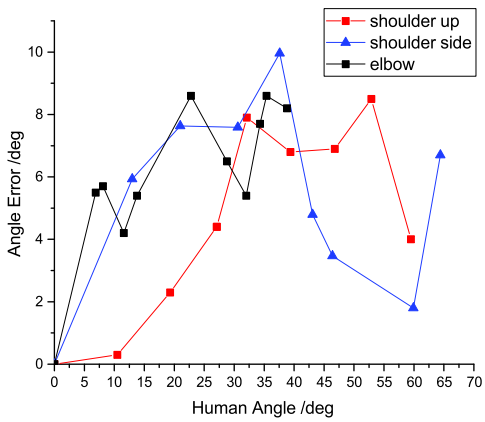


Table II Distance error

Experiment no.	Distance error Xa/mm	Ya/mm	Rb/mm
1	0.0	0.0	0.0
2	2.6	−4.1	4.9
3	−14.3	2.2	14.5
4	−4.8	0.4	4.8
5	−8.4	−2.2	8.7
6	−10.4	−0.5	10.4
7	−13.8	−3.2	14.2
8	−7.3	2.3	7.7
9	−3.0	−1.2	3.2
10	0.0	0.0	0.0
Avgc /mm	−5.9	−0.6	6.8
|Avgc |/mm	6.5	1.5	6.8

Notes: [a] The directions of X and Y are shown in Figure 15(a). [b] R represents the distance error from the expected point and it follows $R = sqrt(X^2 + Y^2)$. [c] Avg is the abbreviation for average values

in the experiment, the demonstrator first teaches the robot to approach the left and right holes of 3 flanges in sequence and then the robot is requested to approach 6 holes continuously and automatically in the following experiments without demonstration. The diameter of these holes is close to that of the fingertip. Figure 15 shows the teaching process and Table III shows the success rates. The experiment is carried out 5 times and failure mainly happens when the finger collides with the flange due to accumulated translational errors. The high success rate can validate the high repeatability and reliability of the proposed natural teaching paradigm. Once the demonstration is completed and the desired position of the end-effector is also reached, the robot can repeat the same motion using the recorded joint angles during the whole process.

5.3 Dexterity

With the proposed natural teaching paradigm, the humanoid robot is capable of accomplishing complicated movements. Two experiments are conducted to demonstrate the dexterity of natural teaching. First, an eye-body synergic experiment is performed via scene-motion cross-modal perception. In the experiment, the robot is confronted with a complicated situation where flanges and other things heap up together on the table. As is the same with the aforementioned teaching experiment, the robot is expected to approach the inner circle of each flange with its index finger. The experimental results are shown in Figures 16-18.

The second one is the classical experiment of obstacle avoidance. As shown in Figure 19, the end-effector needs to cross the obstacle first before it reaches the desired position and orientation. The trajectory is generated by means of natural teaching. In such a complicated mission scene, motion planning always consumes a great amount of computation and time, while natural teaching can fully utilize human's perception and decision-making ability, making it more convenient and less time-consuming. The scene-motion cross-modal perception enables human to perceive the surroundings

Figure 15 Demonstration of approaching six holes

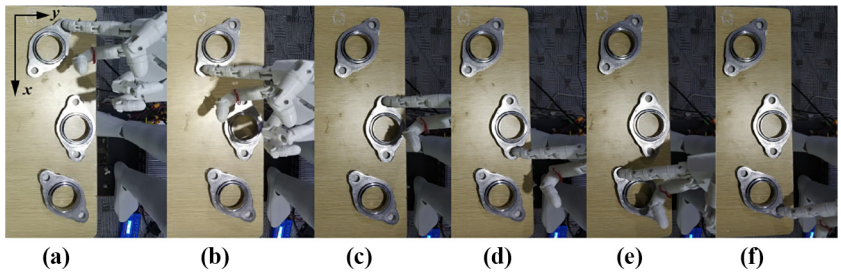


Table III Results of repeating process

Experiment no.	a	b	c	d	e	f
1	S	S	S	S	S	S
2	S	S	S	S	S	S
3	S	S	S	S	S	F
4	S	S	S	F	S	S
5	S	S	F	S	S	S
Success rate (%)	100	100	80	80	100	80

Notes: a, b, c, d, e and f represent each hole in Figure 15. F stands for failure and S stands for success

Figure 16 Approaching an inclined flange

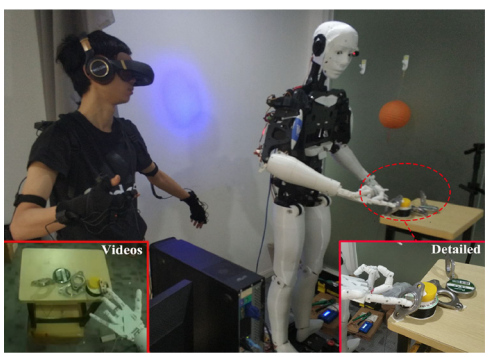


Figure 17 Approaching a vertical flange

Figure 18 Approaching an occluded flange

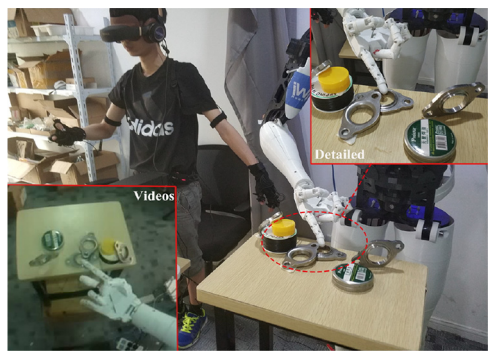
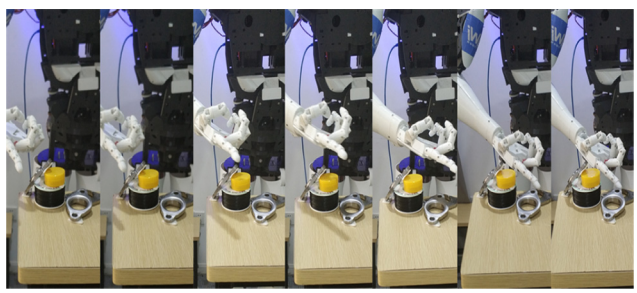


Figure 19 Snapshots of end-effector crossing obstacles



around the robot and the robot to reproduce human's motions. Hence, complicated missions can usually be accomplished with natural teaching at a minimum cost.

6. Conclusions

This paper presents a novel natural teaching paradigm for a humanoid robot. A perception system composed of a vision sensor and a motion capture system realizes the cross-modal combination of scene and motion information. Through multiple visualization terminals for different motion systems, a fast mapping algorithm and reliable data transfer methods, real-time motion imitation is accomplished. Several experiments are designed to validate the accuracy, repeatability, and dexterity of the proposed natural teaching paradigm. Through natural teaching from FPV, human intelligence builds connections between scene information and movement policy, thus making it possible for robots to make autonomous decisions based on

cross-modal perception. Future work will lay more emphasis on the development of the perception system to improve the user experience as well as the accuracy of motion imitation. Encouraged by Tri-Co Robot initiative (Ding *et al.*, 2017), we hope this work will further contribute to the enhancement of industrial robot intelligence.

References

Alibeigi, M., Ahmadabadi, M.N. and Araabi, B.N. (2017), "Fast, robust, and incremental model for learning high-level concepts from human motions by imitation", *IEEE Transactions on Robotics*, Vol. 33 No. 1, pp. 153-168.

Argall, B.D., Chernova, S., Veloso, M. and Browning, B. (2009), "A survey of robot learning from demonstration", *Robotics and Autonomous Systems*, Vol. 57 No. 5, pp. 469-483.

Bindal, A., Kumar, A., Sharma, H. and Kumar, W.K. (2015), "Design and implementation of a shadow bot for mimicking the basic motion of a human leg", *Proceedings of International Conference on Recent Developments in Control, Automation and Power Engineering*, pp. 361-366.

Bohg, J., Morales, A., Asfour, T. and Kragic, D. (2014), "Data-driven grasp synthesis-a survey", *IEEE Transactions on Robotics*, Vol. 30 No. 2, pp. 289-309.

Dai, H., Cai, B., Song, J. and Zhang, D. (2010), "Skeletal animation based on bvh motion data", *Proceedings of 2nd International Conference on Information Engineering and Computer Science*, pp. 1-4.

Deisenroth, M.P., Neumann, G. and Peters, J. (2013), "A survey on policy search for robotics", *Foundations and Trends in Robotics*, Vol. 2 Nos 1/2, pp. 1-142.

Ding, H., Yang, X., Zheng, N., Li, M., Lai, Y. and Wu, H. (2017), "Tri-co robot: a Chinese robotic research initiative for enhanced robot interaction capabilities", *National Science Review*, Vol. 5 No. 6, pp. 799-801.

Ding, I., Chang, C. and He, C. (2014), "A kinect-based gesture command control method for human action imitations of humanoid robots", *Proceedings of 2014 International Conference on Fuzzy Theory and Its Applications*, pp. 208-211.

Durdu, A., Cetin, H. and Komur, H. (2014), "Robot imitation of human arm via artificial neural network", *Proceedings of 16th International Conference on Mechatronics - Mechatronika 2014*, pp. 370-374.

Finn, C. Yu, T. Zhang, T. Abbeel, P. and Levine, S. (2017), "One-shot visual imitation learning via Meta-learning", *arXiv preprint arXiv:1709.04905*.

Gobee, S., Muller, M., Durairajah, V. and Kassoo, R. (2017), "Humanoid robot upper limb control using microsoft kinect", *Proceedings of International Conference on Robotics, Automation and Sciences*, pp. 1-5.

Gong, L., Gong, C., Ma, Z., Zhao, L., Wang, Z., Li, X., Jing, X., Yang, H. and Liu, C. (2017), "Real-time human-in-the-loop remote control for a life-size traffic police robot with multiple augmented reality aided display terminals", *Proceedings of 2nd International Conference on Advanced Robotics and Mechatronics*, pp. 420-425.

Kassahun, Y., Yu, B., Tibebu, A.T., Stoyanov, D., Giannarou, S., Metzen, J.H. and Vander Poorten, E. (2016), "Surgical robotics beyond enhanced dexterity instrumentation: a survey of machine learning techniques and their role in intelligent and autonomous surgical actions", *International Journal of Computer Assisted Radiology and Surgery*, Vol. 11 No. 4, pp. 553-568.

Koenemann, J., Burget, F. and Bennewitz, M. (2014), "Real-time imitation of human whole-body motions by humanoids", *Proceedings of IEEE International Conference on Robotics and Automation*, pp. 2806-2812.

Koenig, N. and Matarić, M.J. (2017), "Robot life-long task learning from human demonstrations: a bayesian approach", *Autonomous Robots*, Vol. 41 No. 5, pp. 1173-1188.

Koert, D., Maeda, G., Lioutikov, R., Neumann, G. and Peters, J. (2016), "Demonstration based trajectory optimization for generalizable robot motions", *Proceedings of IEEE-RAS 16th International Conference on Humanoid Robots*, pp. 515-522.

Langevin, G. (2014), "Inmoov", available at: www.inmoov.fr/project (accessed 6 March 2019).

Laskey, M., Lee, J., Chuck, C., Gealy, D., Hsieh, W., Pokorny, F.T., Dragan, A.D. and Goldberg, K. (2016), "Robot grasping in clutter: using a hierarchy of supervisors for learning from demonstrations", *Proceedings of IEEE International Conference on Automation Science and Engineering*, pp. 827-834.

Levine, S., Pastor, P., Krizhevsky, A., Ibarz, J. and Quillen, D. (2018), "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection", *The International Journal of Robotics Research*, Vol. 37 Nos 4/5, pp. 421-436.

Lim, A. and Okuno, H.G. (2014), "The MEI robot: towards using motherese to develop multimodal emotional intelligence", *IEEE Transactions on Autonomous Mental Development*, Vol. 6 No. 2, pp. 126-138.

Lim, G.H. (2016), "Two-step learning about normal and exceptional human behaviors incorporating patterns and knowledge", *Proceedings of IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pp. 162-167.

Liu, K., Fridman, E., Johansson, K.H. and Xia, Y. (2016), "Quantized control under round-robin communication protocol", *IEEE Transactions on Industrial Electronics*, Vol. 63 No. 7, pp. 4461-4471.

Meng, X., Pan, J. and Qin, H. (2017), "Motion capture and retargeting of fish by monocular camera", *Proceedings of International Conference on Cyberworlds*, pp. 80-87.

Michini, B., Cutler, M. and How, J.P. (2013), "Scalable reward learning from demonstration", *Proceedings of IEEE International Conference on Robotics and Automation*, pp. 303-308.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A. K. and Ostrovski, G. (2015), "Human-level control through deep reinforcement learning", *Nature*, Vol. 518 No. 7540, p. 529.

Noda, K., Arie, H., Suga, Y. and Ogata, T. (2014), "Multimodal integration learning of robot behavior using deep neural networks", *Robotics and Autonomous Systems*, Vol. 62 No. 6, pp. 721-736.

Osentoski, S., Pitzer, B., Crick, C., Jay, G., Dong, S., Grollman, D., Suay, H.B. and Jenkins, O.C. (2012), "Remote robotic laboratories for learning from demonstration", *International Journal of Social Robotics*, Vol. 4 No. 4, pp. 449-461.

Riley, M., Ude, A., Wade, K. and Atkeson, C.G. (2003), "Enabling real-time full-body imitation: a natural way of transferring human movement to humanoids", *Proceedings of IEEE International Conference on Robotics and Automation*, Vol. 2, pp. 2368-2374.

Rodriguez, N.E.N., Carbone, G. and Ceccarelli, M. (2006), "Antropomorphic design and operation of a new low-cost humanoid robot", *Proceedings of 1st IEEE/RAS-EMBS International Conference on Biomedical Robotics and Biomechatronics*, pp. 933-938.

Rozo, L., Calinon, S., Caldwell, D.G., Jiménez, P. and Torras, C. (2016), "Learning physical collaborative robot behaviors from human demonstrations", *IEEE Transactions on Robotics*, Vol. 32 No. 3, pp. 513-527.

Silver, D., Bagnell, J.A. and Stentz, A. (2012), "Active learning from demonstration for robust autonomous navigation", *Proceedings of IEEE International Conference on Robotics and Automation*, pp. 200-207.

Wang, Z., Gong, L., Chen, Q., Li, Y., Liu, C. and Huang, Y. (2016), "Rapid developing the simulation and control systems for a multifunctional autonomous agricultural robot with ROS", *International Conference on Intelligent Robotics and Applications*, pp. 26-39.

Wächter, M. and Asfour, T. (2015), "Hierarchical segmentation of manipulation actions based on object relations and motion characteristics", *Proceedings of International Conference on Advanced Robotics*, pp. 549-556.

Yavşan, E. and Uçar, A. (2016), "Gesture imitation and recognition using kinect sensor and extreme learning machines", *Measurement*, Vol. 94, pp. 852-861.

Corresponding author

Liang Gong can be contacted at: gongliang_mi@sjtu.edu.cn