

# Hand Detection and Location Based on Improved SSD for Space Human-robot Interaction

Qing Gao<sup>1,2</sup>, Jinguo Liu<sup>1</sup>, Zhaojie Ju<sup>3</sup>, Lu Zhang<sup>4</sup>, Yangmin Li<sup>5</sup>

<sup>1</sup> The State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China

liujinguo@sia.cn

<sup>2</sup> University of the Chinese Academy of Science, Beijing 100049, China

<sup>3</sup> School of Computing, University of Portsmouth, Portsmouth, PO1 3HE, U.K

<sup>4</sup> Key Laboratory of Space Utilization, Technology and Engineering Center for space Utilization, Chinese Academy of Sciences, Beijing 100094, China

<sup>5</sup> Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hong Kong, 999077, China

**Abstract.** In the astronaut-space robot interaction using hand gestures, the detection and location of hands are the premise and basis of vision-based hand gesture recognition and hand tracking. In this paper, the SSD (Single Shot Multibox Detector) which is a kind of Convolutional Neural Network (CNN) model is utilized to detect and locate astronaut's hands for space human-robot interaction (SHRI) based on hand gestures. First of all, in order to meet the need of hand detection and location, an improved SSD model is designed to detect hands when they are shown as small targets in images. Then, a platform for SHRI is built and a set of hand gestures for SHRI are designed. Finally, the proposed SSD model is validated experimentally on a homemade hand gesture database for proving the superiority of this improved SSD model to small target hands detection.

**Keywords:** Human-robot Interaction, Hands Detection, SSD, CNN.

## 1 Introduction

Nowadays, in the space missions, space robots are often controlled interactively by staff or astronauts because of the limited intelligence of space robots and safety. In general, the advantage of astronaut is that it has a strong sense of perception, decision-making and planning capabilities, while the advantage of the space robot is that it can achieve smooth, high-precision, wide-range operations. SHRI technology effectively combines the advantages of astronauts and space robots and plays an important role in space missions [1]. Among them, the hand gesture-based SHRI with its natural and intuitive, informative, non-contact advantages is very suitable for applications in SHRI tasks. At present, there have been some achievements in this field at home and abroad [2, 3, 4]. At the same time, the design of SHRI system and the design of interactive hand gestures are crucial to gesture-based SHRI. A set of reasonable and natural SHRI hand gestures can help astronauts to control space robots efficiently and conveniently.

Astronauts' hands detection and location are premise and basis of hand gestures recognition and hands tracking. They play a key role in the entire hand gesture interaction process. At present, significant efforts have been made in the field of hands detection and location [5, 6, 7]. Among them, deep learning (DL) has made breakthrough progress in visual-based hand gestures interaction [8, 9, 10, 11]. For the target detection and location problems, the related deep learning models have R-CNN, Fast R-CNN, Faster R-CNN, YOLO and SSD. Among them, the SSD model has excellent performance in the real-time detection and location of targets. It uses end-to-end training method to achieve a balance between speed and accuracy (More accurate than Faster R-CNN and faster than YOLO) [12]. However, the SSD model also has the problem of inaccurate detection of small targets. For example, if the astronaut is far away from the camera during operation, the hand may be considered as a small target. In this situation, the SSD model can't detect the hand accurately. Therefore, the SSD model needs to be improved to adapt to the detection and location of the hand when it considered as a small target.

In this paper, aiming at hand detection and location problems, the SSD model is improved. A feature-based SSD model is designed to detect the hand precisely when it is a small target. In addition, aiming at SHRI tasks, the second-generation astronaut assistant robot (AAR-2) which was developed by Shenyang Institute of Automation (SIA), Chinese Academy of Sciences (CAS) is chose to act as the space robot to set up the SHRI system platform. What's more, a set of hands-coordination interactive hand gestures which are called "left hand for instruction and right hand for operation" is designed. In the experiment part, a set of interactive astronaut hand gestures database is manufactured by ourselves. Under this database, the SSD model and the improved SSD model in this paper are trained and verified respectively, and their test results are compared.

The rest of this paper is structured as follows: In section 2 the improved SSD model will be introduced. In section 3 the SHRI system and the hands-coordination interactive hand gestures will be introduced. And in section 4 the SSD model and the improved SSD model will be tested on a homemade astronaut hand gestures interactive database, and the experiment results will be compared to analyze the effectiveness and superiority of the improved SSD model. At last, the conclusion and the prospect of this article will be shown in section 5.

## **2 Hand Detection and Location Method Based on Improved SSD Model**

Based on the SHRI, the requirements for astronaut's hand detection and location are shown as follows:

- (1) Ensure the real-time performance of hand detection and location;
- (2) Can detect and locate different hand gestures;
- (3) When the astronaut's hand is shown as a small target in the image, it can also be detected and located precisely.

## 2.1 Introduction of SSD Model

Based on the above requirements, using SSD model to detect and locate the astronaut's hand can learn the characteristics of different hand gestures and can guarantee the real-time performance. The network structure is shown in Fig.1.

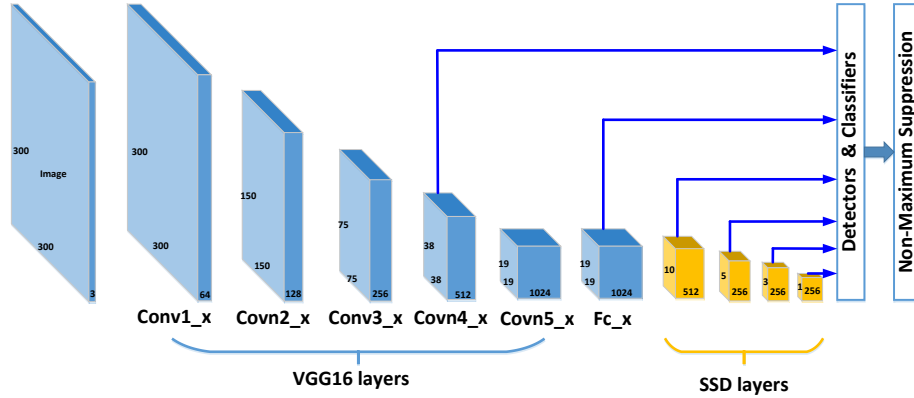


Fig. 1. SSD architecture.

As shown in Fig.1, the SSD is a full convolution neural network detector with different layers to detect objects of different sizes. However, the detection of small targets by the SSD model often does not perform well. The reason is that in shallower layers, the feature maps are large with more contextual information, but semantic information is not enough; in deeper layers, the semantic information is much enough, but after too many pooling layers, the feature maps are very small. When the hand to be detected is a small target, it needs a feature map that is large enough to provide finer features, as well as has sufficient semantic information to distinguish hand from the background.

## 2.2 Introduction of Improved SSD Model

In order to solve the above problem, the information of the shallower layer can be combined with the information of the deeper layer to design a layer that has both enough contextual information and enough semantic information. Inspired by DSSD [13] and Feature-Fused SSD [14], the improved SSD model is designed as Fig. 2:

According to the principle of SSD model, shallower layers networks are used to detect small targets, and deeper layers networks are used to detect large targets. As shown in Fig.2, aiming at the insufficient semantic information of shallower layers networks, the shallower layers should be fused with the deeper layers to increase the semantic information of shallower layers networks so as to increase the detection accuracy of the network to small targets. From [14], it can be seen that Conv4\_3 has the best detection effect on small targets. Therefore, in this paper, Conv4\_3 and Conv6\_2 are utilized to fusion according to their features to improve the network detection of small target hand.

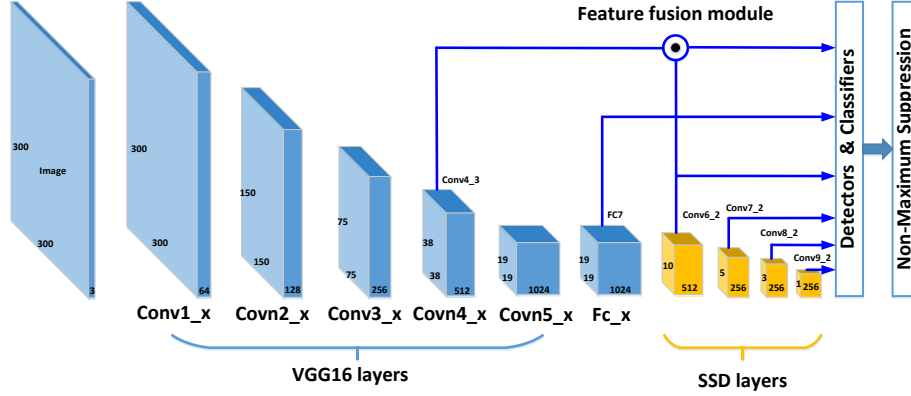


Fig. 2. Improved SSD architecture.

### 2.3 Feature Fusion Module

The feature fusion module is indicated by Fig.3. The specific method is that conduct twice deconvolution operation on the feature map of Conv6\_2 layer to get Deconv6\_2 layer which has the same size as Conv4\_3 layer. Then two  $3 \times 3$  convolutional layers are used after Conv4\_3 layer and Deconv6\_2 layer for learning the better features to fuse. After this, normalization layers are following with different scales respectively, i.e. 10, 20. Finally, fuse the feature maps of the two layers according to element-wise product method to produce a new feature map Fusion\_conv43\_conv62 to detect and locate small targets (From [13], it can be seen that element-wise product method can get the best accuracy result).

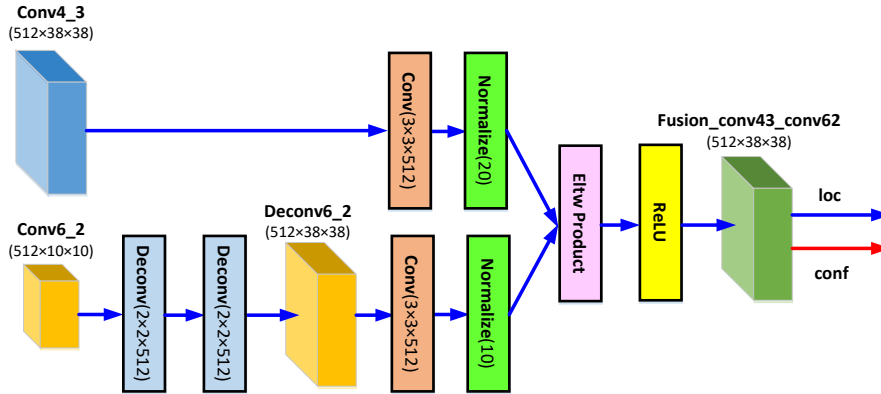


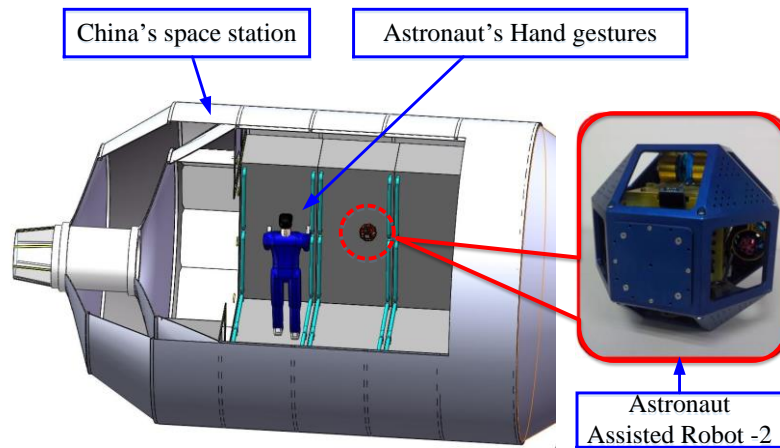
Fig. 3. Feature fusion module.

### 3 SHRI System

#### 3.1 Experiment Platform

The SHRI system includes three parts which are space robot system, astronaut hand gestures operation system and hand gestures recognition system. Astronaut issues control commands through specific hand gestures, then the images containing hand gestures are transmitted to the hand gestures recognition system via an image capture device. After this, the hand gestures recognition system can detect, locate and track the hands and recognize the semantics of hand gestures. Finally, hand gestures can be decoded into control signals and then send them to the space robot to control its operation. According to the multi-spatial scope between astronaut and space robot, the SHRI can be divided into astronaut-robot shared space (shoulder-to-shoulder) and astronaut-robot non-shared space (Line-of-sight, Over-the-horizon, Interplanetary) [1]. In the astronaut-robot shared space mode, the image capture device and the gesture recognition system can be mounted on the space robot. While in the astronaut-robot non-shared space mode, the image capture device and gesture recognition system need to be installed on the console.

Select the AAR-2 as the space robot [15, 16]. This robot works at the space station and is utilized to assist astronauts to complete some space missions. It has six degrees of freedom, and can fly freely in the space station cabin. So its movement is very suitable for hand gestures interactive control. The schematic diagram that the astronaut uses hand gestures to control the AAR-2 is shown as Fig.4.



**Fig. 4.** The schematic diagram that the astronaut uses hand gestures to control the AAR-2.

Using the FT-200 miniature simulation air-floating platform to simulate the space micro-gravity environment. The AAR robot can be mounted on it and can move on a marble platform with three degrees of freedom motion (translate along X axis and Y axis and rotate around z axis). The Kinect is selected as the image capture device to

capture astronaut's hand gestures information. In terms of software, at the astronaut's hand gestures images recognition part, choose ROS (Robot Operating System) as the operating system; at the Deep Learning part, choose Caffe (Convolutional Architecture for Fast Feature Embedding) as the framework for deep learning models; at robot system part, the STM32 processor is selected as the control unit for controlling robot drivers and related motion control algorithm. The whole SHRI system experiment platform is illustrated in Fig.5.

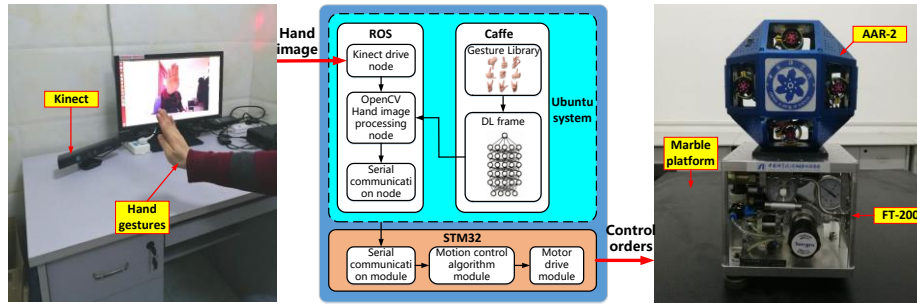


Fig. 5. SHRI system experiment platform.

### 3.2 Hand Gestures For SHRI









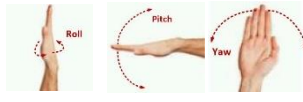


There are two types of hand gestures in SHRI: one is command-type hand gestures. It means that astronauts communicate certain deterministic semantic information to space robots through changing the morphology or spatial orientation of their hands. The other is control-type hand gestures. It means that astronauts transfer quantitative parameters to space robots by moving their hands.

Based on the above, a set of hands-coordination interactive hand gestures which are called "left hand for instruction and right hand for operation" is designed based on the ASL hand gestures [17]. The specific hand gestures semantic correspondence is shown as Tab.1: astronaut can use the left hand to set up 8 semantics commands those are "Begin to control", "Stop control", "Finish control", "Path tracking", "Linear motion", "Rotational Motion", "Object approximation" and "Data transmission". When the left hand is recognized as the "Begin to control" gesture command, the robot performs a response to start control. After starting the control, when the left hand is recognized as the "Path tracking", the space robot will follow the track of right hand; when the left hand gesture is recognized as the "Linear motion", the space robot will make translate with right hand; and when the left hand gesture is recognized as the "Rotational Motion", the space robot will rotate with the right hand. when the left hand gesture is recognized as the "Object approximation", the space robot will move to the target; And when the left hand gesture is recognized as the "Data transmission", the robot will send position and attitude data and sensor test data to the host computer; Each time a hand gesture is completed, the left hand can be transformed into a "Finish control" gesture, then it will switch to a next gesture command. When the left hand is recognized as the "Stop control", interactive control between astronaut and space robot will end, and the gesture

commands are no longer valid unless the next "Begin to control" hand gesture is detected.

As can be seen from Table 1, recognizing the left hand instructions requires the recognition of hand gesture semantics [18]. While recognizing the right hand operation requires tracking the hand in six degrees of freedom. And hand detection and location are crucial in both hand gestures semantic recognition and hand tracking. Therefore, detection and location of astronaut's hands are researched deeply in the following content.

**Table 1.** Chart of astronaut-space robot hands-coordination hand gestures.

Hand gesture semantics	ASL letters	Left hand gestures	Right hand gestures
Begin to control	<i>B</i>		-
Stop control	<i>S</i>		-
Finish control	<i>F</i>		-
Path tracking	<i>P</i>		
Linear motion	<i>L</i>		
Rotational motion	<i>R</i>		
Object approximation	<i>O</i>		-
Data transmission	<i>D</i>		-

## 4 Experiments

### 4.1 Hand Gestures Database For SHRI

The Hand gestures database for SHRI need to meet the following requirements:

- (1) Have the 8 hand gestures of the chart of astronaut-space robot hands-coordination hand gestures;
- (2) Have images contain hands of different sizes.

Since the known hand gestures databases do not meet all the above requirements, a set of Space Robot Simple Sign Language (SRSSL) database was made by ourselves. This hand gestures database includes six different human hand gestures RGB images. Each of them contains the eight types hand gestures in the chart of astronaut-space robot hands-coordination hand gestures, and includes five different sized hand gestures. Each person has 1000 hand gestures images, so the database contains a total of 6,000 hand gestures images. Select five people's hand gestures images (5000 images) as the training data, and another person's hand gestures images (1000 images) as the test data. A part of images ("Begin to control" hand gestures images) of SRSSL database are shown in Fig.6. Where  $PN$  denotes the person number,  $XS$ ,  $S$ ,  $M$ ,  $L$ ,  $XL$  denote five different sizes hand targets. For example,  $XS$  denotes the minimum hand target and  $XL$  denotes the maximum hand target.

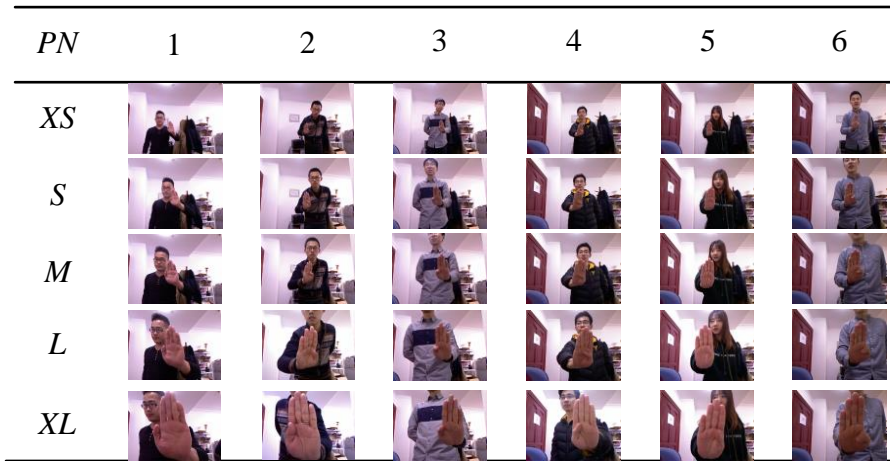


Fig. 6. SRSSL database.

Use the above SRSSL database to evaluate the proposed improved SSD model. The hand detection and location experiment will be conducted and the experiment results will be compared with the results of the SSD model.

#### 4.2 Hand Detection And Location Experiment For Small Target Hands

The hardware equipment of these experiments is shown as follows: Intel Core I5-6400 CPU, NVIDIA GeForce GTX 1060 6GB GDDR5, 16GB ROM. And the experiments are conducted in Caffe environment under Ubuntu 14.04 64bit OS system.

This article is primarily to solve the problem that the SSD model can't detect small targets precisely. In order to validate the improved SSD model proposed in this paper, the SSD model and the improved SSD model are trained and tested on the above SRSSL database respectively. The types of recognition include the 8 kinds of SHRI hand gestures.

Table 2 shows the average accuracy of the SSD method and the improved SSD method and the accuracy of each type of hand gesture.

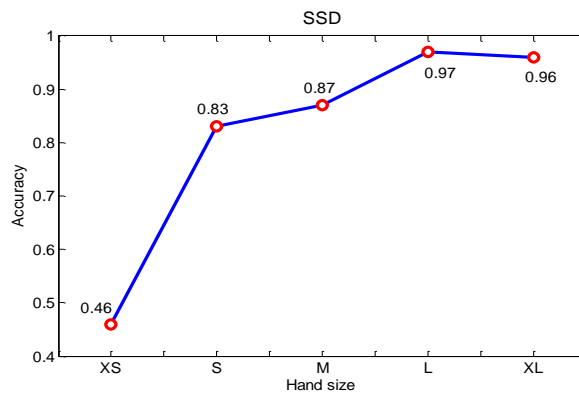
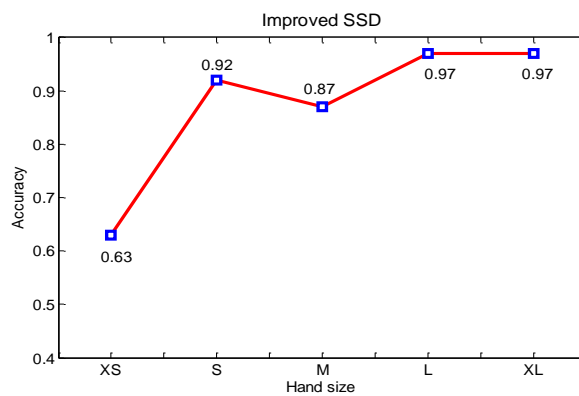


**Table 2.** The comparison of detection results on SRSSL database.

Methods	<i>B</i>	<i>S</i>	<i>F</i>	<i>P</i>	<i>L</i>	<i>R</i>	<i>O</i>	<i>D</i>	mAP
SSD	79.1	83.5	75.6	70.4	86.7	85.8	86.8	86.9	81.8
Improved SSD	86.9	89.5	83.5	78.1	89.1	89.4	90.3	90.8	87.2

From Table 2, it can be seen that the mAP of improved SSD proposed in this paper is 87.2%. Compared with the SSD method, the accuracy improves 5.4%. What's more, the detection accuracy of each hand gesture of improved SSD is higher than SSD. Therefore, the improved SSD in this paper can contribute to improving the detection accuracy of hand gestures.

In order to validate the accuracy of the improved SSD to the recognition of small target hand, the SSD and improved SSD are used to experiment on five different sizes of hand images respectively. The detection accuracies of hands in different sizes are shown as follows.

**Fig. 7.** The hand detection accuracies of SSD.**Fig. 8.** The hand detection accuracies of improved SSD.

Comparing the data shown in Fig.7 and Fig.8, it can be seen that when the hand sizes are *XS* and *S*, the detection accuracies of the SSD model are 46% and 83% respectively. While the detection accuracies of the improved SSD model are 63% and 92% respectively. And when the hand sizes are *M*, *L* and *XL*, the detection accuracy of improved SSD model are similar to that of SSD model. So the improved SSD model makes a significant improvement on the precision of hand detection of small targets compare with the SSD model. And the detection accuracy of large target hand is also high.

For a same RGB image with a small target hand, the detection and location effect pictures of SSD and improved SSD are shown in Fig.9.



**Fig. 9.** The left picture is the detection and location effect picture of SSD, and the right picture is the detection and location effect picture of improved SSD.

As can be observed in Fig.9, the red boxes in the pictures are the detected gestures of networks. The blue words are the credibility of hands. Although both models can detect the hand successfully when it is displayed as a small target, but the improved SSD model predicts a higher credibility (66.4082%) than SSD model (50.1292%). Therefore.

In a word, the experiments fully demonstrate the effectiveness of the improved SSD model for small target hand detection and location.

## 5 Conclusion

In this paper, we mainly dealt with hand gestures interaction tasks in SHRI and researched on hands detection and location. Where the innovations of this paper includes: (a) Aiming at the problem that SSD model does not work well for small target detection, an improved SSD model was proposed. It fused the features of Conv4\_3 layer and Conv6\_2 layer to improve the detection of small target hands; (b) In the SHRI system, a set of hands-coordination interactive hand gestures which is called “left-handed instruction and right-handed operation” was designed; (c) Homemade a set of SRSSL database for experiments of hands detection and location.

The future work mainly includes: (a) Distinguishing and locating the right hand and the right hand and ensuring the real-time performance of the hand detection and location; (b) For the "right-handed operation", the six-DOF tracking of hand will be researched.

**Acknowledgments.** This work is supported by Research Fund of China Manned Space Engineering (050102), the Key Research Program of the Chinese Academy of Sciences (Y4A3210301), the National Science Foundation of China (51175494, 61128008, 51575412 and 51775541).

## References

1. Fong, T., Nourbakhsh, I.: Interaction Challenges in Human-Robot Space Exploration. *Interactions*. 12(2), 42–45 (2005).
2. <http://www.pingwest.com/leap-motion-meets-nasa/>, last accessed 2018/01/31.
3. Wolf, M.T., Assad, C., Vernacchia, M.T. et al. Gesture-Based Robot Control with Variable Autonomy from the JPL BioSleeve. In: *Robotics and Automation (ICRA)*, 2013 IEEE International Conference, pp. 1160-1165. IEEE, Karlsruhe, Germany (2013).
4. Liu, J.G., Luo, Y.F., Ju, Z.J.: An Interactive Astronaut-Robot System with Gesture Control. *Computational Intelligence and Neuroscience*. 2016 (2016).
5. Grzejszczak, T., Kawulok, M., Galuszka, A.: Hand landmarks detection and localization in color images. *Multimedia Tools and Applications*. 75(23), 16363–16387 (2016).
6. Grzejszczak, T., Łęgowski, A., Niezabitowski, M.: Application of hand detection algorithm in robot control. In: *17th International Carpathian Control Conference (ICCC)*. IEEE, Tatranska Lomnica, Slovakia (2016).
7. Raheja, J.L., Chaudhary, A., Maheshwari, S.: Hand gesture pointing location detection. *International Journal for Light and Electron Optics*. 125(3), 993-996 (2014).
8. Tompson, J., Stein, M., Lecun, Y., Perlin, K.: Real-Time Continuous Pose Recovery of Human Hands Using Convolutional Networks. 33(5) (2014)
9. Ge, L., Liang, H., Yuan, J., Thalmann, D.: Robust 3D Hand Pose Estimation in Single Depth Images: from Single-View CNN to Multi-View CNNs. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3593-3601. IEEE, Seattle, WA (2016).
10. Yamashita, T., Watasue, T.: Hand posture recognition based on bottom-up structured deep convolutional neural network with curriculum learning. In: *2014 IEEE International Conference on Image Processing (ICIP)*, pp. 853-857. IEEE, Paris, France (2014).
11. Molchanov, P., Gupta, S., Kim, K.: Hand Gesture Recognition with 3D Convolutional Neural Networks. In: *CVPR 2015*, IEEE, Boston, America (2015).
12. Liu, W., Anguelov, D., Erhan, D., Szegedy, C.: SSD: Single Shot MultiBox Detector. In: *14th European Conference on Computer Vision (ECCV)*, pp. 21-37. IEEE, Amsterdam, The Netherlands (2016).
13. Fu, C.Y., Liu, W., Ranga, A., Tyagi, A., Berg, A.C.: DSSD : Deconvolutional Single Shot Detector. *Computer Vision and Pattern Recognition*. arXiv:1701.06659 (2017).
14. Cao, G.M., Xie, X.M., Yang, W.Z., et al.: Feature-Fused SSD: Fast Detection for Small Objects. *Computer Vision and Pattern Recognition*. arXiv:1709.05054 (2017).
15. Liu, J.G., Gao, Q., Liu, Z.W., Li, Y.M.: Attitude Control for Astronaut Assisted Robot in the Space Station. *International Journal of Control, Automation and Systems*. 14(4), 1082-1095 (2016).

16. Gao, Q., Liu, J.G., Tian, T.T., Li, Y.M.: Free-flying dynamics and control of an astronaut assistant robot based on fuzzy sliding mode algorithm. *Acta Astronautica*. 138, 462-474 (2017).
17. Gattupalli, S., Ghaderi, A., Athitsos, V.: Evaluation of Deep Learning based Pose Estimation for Sign Language Recognition. In: 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments. IEEE, Greece (2016).
18. Gao, Q., Liu, J.G., Ju, Z.J., et al.: Static hand gesture recognition with parallel CNNs for space human-robot interaction. In: 9th International Conference on Intelligent Robotics and Applications (ICIRA), pp. 462-473. Springer, Wuhan, China (2017).