Robotica (2022), **40**, pp. 365–376
doi:10.1017/S0263574721000618

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Assist system for remote manipulation of electric drills by the robot "WAREC-1R" using deep reinforcement learning

Xiao Sun[1*], Hiroshi Naito[2], Akio Namiki[3], Yang Liu[3], Takashi Matsuzawa[2], and Atsuo Takanishi[2]

[1]Department of Mechatronics, University of Yamanashi, Yamanashi, Japan, [2]Department of Modern Mechanical Engineering, Waseda University, Tokyo, Japan and [3]Department of Mechanical Engineering, Chiba University, Chiba, Japan
*Corresponding author. Email: xsun@yamanashi.ac.jp

Received: 27 June 2020; **Revised:** 29 April 2021; **Accepted:** 30 April 2021; **First published online:** 4 June 2021

Keywords: Manipulation, Deep reinforcement learning, Remote operation

Abstract

Operation of tools has long been studied in robotics. Although appropriate hold of the tool by robots is the base of successful tool operation, it is not with ease especially for tools with complicated shape. In this paper, an assist system for a four-limbed robot is proposed for remote operation of reaching and grasping electric drills using deep reinforcement learning. Through comparative evaluation experiments, the increase of success rate for reaching and grasping is verified and the decrease in both physical and mental workload of the operator is also validated by the index of NASA-TLX.

1. Introduction

Manipulation by robots has been highly concerned since the birth of robots, and it remains to be one of the most valuable robotic applications that people have been benefitted from. Operation of tools by robots is not an exception, either. Robots capable of carrying and operating tools themselves without the help of human could be very powerful for various situations. To realize such a goal, the authors developed the four-limbed robot "WAREC-1R" [1] under the fund and support of the project "ImPACT Tough Robotics Challenge" [2] initiated by the Cabinet Office of Japan. This project aims at developing five types of tough robot working in extreme environment (especially for disaster sites), and WAREC-1R is one of them. Figure 1 presents the summary of all locomotion styles WAREC-1R is capable of and an expected application scenario. The long-term goal of WAREC-1R is utilizing different locomotion styles to reach an arbitrary target position in real extreme environment and performing manipulation tasks with high difficulty like the operation of electric tools.

So far, WAREC-1R has realized multiple locomotion styles [3, 4] and verified its high performance and versatility in locomotion. In terms of manipulation, end effector capable of manipulating common target objects in infrastructures such as buttons, switches, and levers was developed [5] with compatibility of all locomotion styles that WAREC-1R is capable of. It also enables remote operation of tools like electric drills with a trigger in case the operation task cannot be completed autonomously. Considering the case that WAREC-1R is stopped by a wall or any obstacles, in our opinion, it would be a good idea to open holes on it/them by drills so that other types of robot with smaller diameters (like snake robots) could pass through the holes and continue performing tasks. Therefore, electric drills are mainly discussed in this paper as the target of operation.

However, remote operation of WAREC-1R has been suffering from the issue that the operation of tools highly depends on the successful grasp, while success rate of grasp was too low to be practical even in remote operation, and eventually most of the grasps were assisted by human. Main reasons can be concluded as (a) design of end effector [5] and (b) insufficient assist/information provided to the

© The Author(s), 2021. Published by Cambridge University Press.

https://doi.org/10.1017/S0263574721000618 Published online by Cambridge University Press

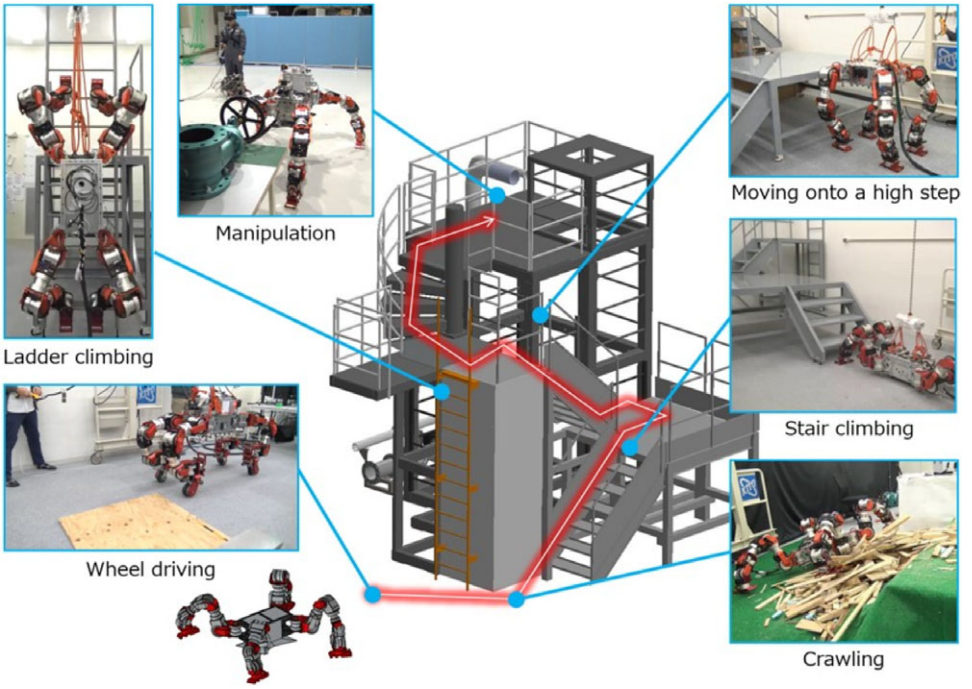


Figure 1. *Achieved locomotion styles and manipulation tasks for WAREC-1R.*

operator. For (a), due to the design of end effector with consideration for both locomotion and manipulation, it has comparatively narrow range of clearance for position and orientation for successful grasp of target like electric drills with complicated shapes. For (b), operators complained that visual information shown on Head Mount Display that operators equipped alone was not enough to determine the proper position (especially for depth) and orientation for successful grasp as well as physical and mental fatigue.

For the robot platform of WAREC-1R that is designed for both locomotion and manipulation, apparently it is not realistic for it to keep holding tools all the way until it arrives the destination and starts manipulation. Thus, grasp (a) with success rate high enough for application and (b) is easy to operate with less burden for operators are highly needed for remote operation of WAREC-1R and its application in tool operation. As the solution, in this paper, we propose an assist system using Deep Reinforcement Learning (DRL in short for the following contents), a subtype of reinforcement learning [6, 7, 8, 9, 10] for both reaching and grasping process before the end effector of WAREC-1R holds the target tool. The contribution and originality of this paper focus on the following points:

(i) Application of DRL to manipulation of tools in reality. As is well known, in recent years, machine learning has been widely applied to robot manipulation (especially for dexterous manipulation) with high performance, with representative instances of refs. [11, 12, 13]. However, few of them realized grasp of tools for operation with complicated shapes like electric drills in reality.

(ii) Furthermore, former research of manipulation using DRL or any other approaches of machine learning tend to either simple grippers that can only hold the target or robot hands with complicated mechatronic structure for end effector, while for the case of WAREC-1R neither gripper nor robot hand are within the scope because of its design concept that guarantees compatibility of both locomotion and manipulation. In fact, end effector of WAREC-1R has much fewer Degree of Freedom (DoF) (three in total) than normal robot hand with fingers and has much more strict constraints in grasping a tool for operation.

https://doi.org/10.1017/S0263574721000618 Published online by Cambridge University Press

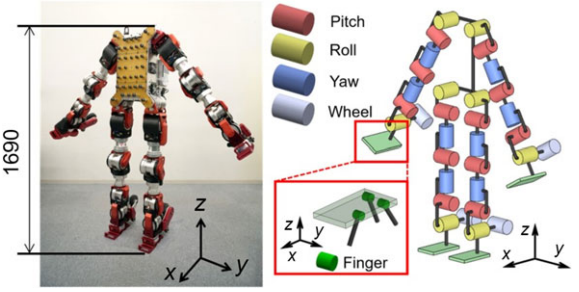


Figure 2. *Overview and DoF configuration of WAREC-1R.*

(iii) DRL application in remote operation for assist of operators' manipulation is rare as well, since tactile information and depth information from force sensors and perception sensors also improve the performance of remote operation, while the approach proposed in this paper with DRL only requires a normal web camera, which greatly simplifies the sensor system of robot and decrease the total cost of robot system as well.

This paper is organized as follows: Section 2 briefly introduces the robot platform "WAREC-1R" as well as hardware components related to manipulation as prerequisite knowledge. Specifically, the end effector of robot for manipulation and devices for remote operation are included besides the robot itself. Section 3 describes the reaching assist using Deep Q Network, an approach of DRL to ensure appropriate approaching to the target tool, which is apparently a necessary process for successful grasp of the tool. Section 4 explains grasping assist with object detection algorithm of YOLO v3 [14], which enables the grasp of target tool by the end effector in appropriate position and orientation. Section 5 presents evaluation experiments of reaching and grasping of an electric drill by remote operation of WAREC-1R, with the comparison between the cases with and without proposed assist system to evaluate its effect. Section 6 draws the conclusion of the whole paper as well as the limit of the proposed assist system, which leads to our prospective work. Finally, in the appendix the mechanism and algorithms of Deep Q Network, an approach used in this paper, are also presented for reference.

2. Robot "WAREC-1R" and its hardware related to manipulation

Before getting into the major part of this study, the hardware systems related to the robot and this study will be introduced briefly. As the improved version of the robot WAREC-1 [15], WAseda REsCuer-No.1 Refined (WAREC-1R) is the human-sized four-limbed robot developed by Takanishi Laboratory, Waseda University, Japan. Its overview, scale, and configuration of Degree of Freedom (DoF) are shown in Fig. 2.

Besides, a newly developed end effector is equipped at the end of its right arm, enabling both locomotion and manipulation. The details of design concept of the end effector can be seen in ref. [5] and its overview as well as details are shown in Fig. 3. The end effector has three fingers for manipulation, in which two opposing fingers are for holding and the one finger left is for operating buttons, switches, triggers, and other small parts on tools. The figure on the right of Fig. 3 shows how the end effector holds an electric drill and operates the drill by pulling the trigger with one finger.

Regarding manipulation, remote operation is available for WAREC-1R with device of Flexible Sensor Tube (FST) and Flexible Sensor Glove (FSG) [16, 17] developed by Namiki Lab in Chiba University, Japan. In addition, Head Mount Display (HMD) is also used for presenting 3D visual information from the stereo camera equipped on the body of robot for operator's operation. FST, FSG, and the operator equipped with them besides HMD are all shown in Fig. 4.

https://doi.org/10.1017/S0263574721000618 Published online by Cambridge University Press

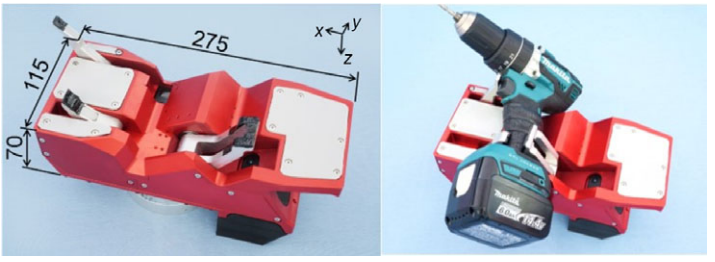


Figure 3. *Overview of the end effector for manipulation with scales.*

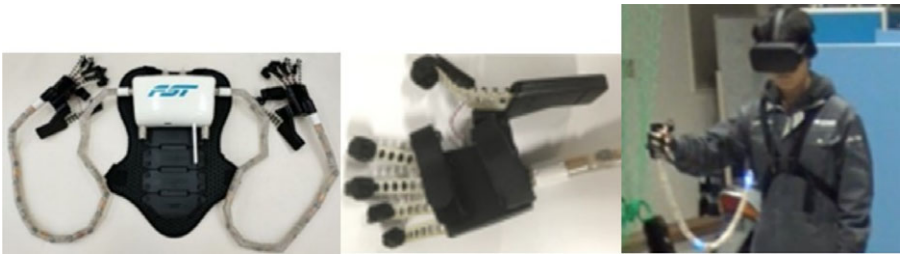


Figure 4. *FST, FSG, and operator equipped with them.*

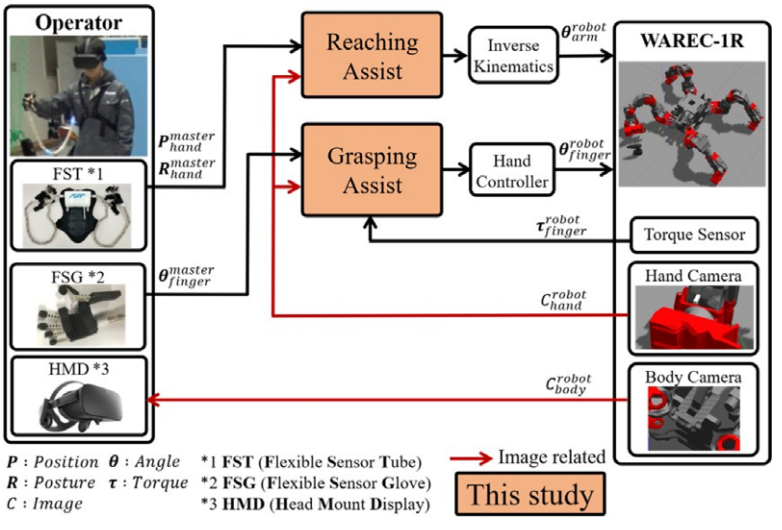


Figure 5. *System configuration of the proposed assist system.*

3. Reaching assist

As is described in ref. [5], end effector in Fig. 3 has been verified to be capable of manipulating buttons, valves, and keys. However, just as Fig. 3 shows successful and stable hold of the tool requires appropriate relative position as well as orientation between the end effector and tool due to the design concept of end effector that it must be sufficiently compact for both locomotion and manipulation. In fact, grasp of the tool in remote operation has been an issue for its low success rate.

Consequently, we propose an assist system to realize stable grasp of the tools (electric drills in this paper) with high success rate. The total system configuration is shown in Fig. 5. In this paper, assist of remote operation acts in two processes: reaching process and grasping process and the former is described in this section.

https://doi.org/10.1017/S0263574721000618 Published online by Cambridge University Press

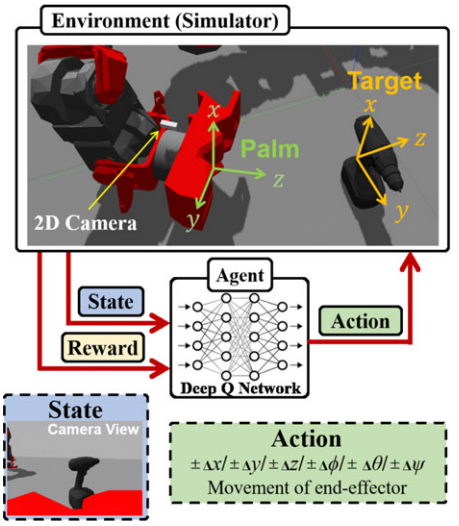


Figure 6. *Training of reaching with Deep Q Network.*

As we can imagine to realize the grasp of tool the end effector must reach, the tool first so that it appears at both appropriate position and orientation for grasping. However, pure manual reaching has been verified to be neither stable nor efficient. Considering that (a) appearance and shape of electric drills differ from each other and (b) most electric drills have common parts: a grip for palm to handle and a trigger for operation, in reaching process Deep Q Network (DQN) [18], an approach of DRL proposed by Google DeepMind, is applied so that reaching process can be used for different types of electric drills that have a grip and a trigger. Figure 6 shows the training and overview of DQN for reaching process. Note that training is done in simulator for efficiency, since training in reality by WAREC-1R takes much longer time and much more effort than training in simulator. Meanwhile, position and orientation of the drill are set randomly within a certain range for better learning.

In the training, DQN deals with the "State", which is the view of camera equipped at the end effector and "Reward" defined by (1)–(4), where the detailed variable definition is given in Table I. The output is "Action", which in this paper is the motion patterns of the end effector, with 12 possible options: moving along $\pm x$, $\pm y$, $\pm z$-axis or rotating around $\pm x$, $\pm y$, and $\pm z$-axis. Here, end effector coordinate system is used with origin defined as a point on the end effector (green coordinate system in Fig. 6) and $+z$-axis defined as the line starting from the origin and pointing at the reference point in the target drill. Thus, this coordinate system changes if the end effector moves. Note that the assist does not work in z-axis because in real application, it is the operator who decides the movement of the end effector in z-axis, which is also the direction of reaching by remote operation.

$$R_{total} = R_{dist} + R_{rot} + R_{as} \tag{1}$$

$$R_{dist} = K_{dist}\{(x_{prev} - x_{cur}) + (y_{prev} - y_{cur}) + (z_{prev} - z_{cur})\} \tag{2}$$

$$R_{rot} = K_{rot}\{(\phi_{prev} - \phi_{cur}) + (\theta_{prev} - \theta_{cur}) + (\psi_{prev} - \psi_{cur})\} \tag{3}$$

$$R_{as} = \begin{cases} C_{as}, & P_{as,i} \leq T_i, i = 1, 2, ..., 5 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Now that training of reaching is done in simulation instead of reality, the difference between simulation and reality may cause the failure of reaching and it is exactly the case for WAREC-1R. Figure 7 shows the inconsistency between camera view in simulator and reality. As arrows point out, reflection on the end effector was not considered at first in the simulation, which is a big issue in realizing successful reaching process. Furthermore, it is also verified that the results of reaching may also be influenced by

https://doi.org/10.1017/S0263574721000618 Published online by Cambridge University Press

Table I. *Definition of variables in training of reaching.*

Variable (vectors in end effector coordinate)	Definition
R_{total}	Total reward
R_{dist}	Reward for the change of position
R_{rot}	Reward for the change of posture
R_{as}	Reward considering assist
K_{dist}	Coefficient for the change of position
K_{rot}	Coefficient for the change of posture
$(x_{cur}, y_{cur}, z_{cur}, \phi_{cur}, \theta_{cur}, \psi_{cur})$	Current drill position and posture
$(x_{prev}, y_{prev}, z_{prev}, \phi_{prev}, \theta_{prev}, \psi_{prev})$	Previous drill position and posture
C_{as}	Constant reward value
$(P_{as,1}, P_{as,2}, ..., P_{as,5}) = (x_{cur}, y_{cur}, \phi_{cur}, \theta_{cur}, \psi_{cur})$	Position (no z_{cur}) and posture for assist
$(T_1, T_2, ..., T_5) = (T_{dist}, T_{dist}, T_{rot}, T_{rot}, T_{rot})$	Threshold of Reward for assist



Figure 7. *Inconsistency of camera view between simulation and reality.*

light condition as well as complexity of background. Therefore, as the solution we added (a) a light with random color and position, (b) trees below the light to create random shadow as well as simulation of reflection, and (c) random backgrounds such as walls and buildings to the training of reaching process shown in Fig. 8.

4. Grasping assist

Due to hardware design concept that enables both locomotion and manipulation as well as the actual size and shape of the end effector, appropriate orientation must be guaranteed for successful grasp after reaching process.

For the judgement of graspability, in this paper, object detection algorithm YOLO v3 is applied. Specifically, photos of the tool in appropriate position and orientation for grasping are taken for the learning under different conditions of lighting and background for robustness, and some of them are shown in Fig. 9 along with training results. One thing worth mentioning is that two different drills are used in this study, and training with one of them works as well for the other one.

Just as Fig. 5 depicts, this block will check whether (1) the robot hand is in the propriate position and orientation to grasp the drill by YOLO v3, (2) there is no excessive torque on fingers for safety, and (3) the operator is closing his hand equipped with FST and FSG, which is the command of grasping. The robot hand will only start grasping the drill if all these three conditions are satisfied.

5. Real-world experiments

As the evaluation of the developed assist system for reaching and grasping of electric drills, we made comparison experiments and the overview can be seen in Fig. 10. Specifically, we asked 21 subjects (14 males and 7 females) to grasp the electric drill randomly placed for the case with and without assist

https://doi.org/10.1017/S0263574721000618 Published online by Cambridge University Press

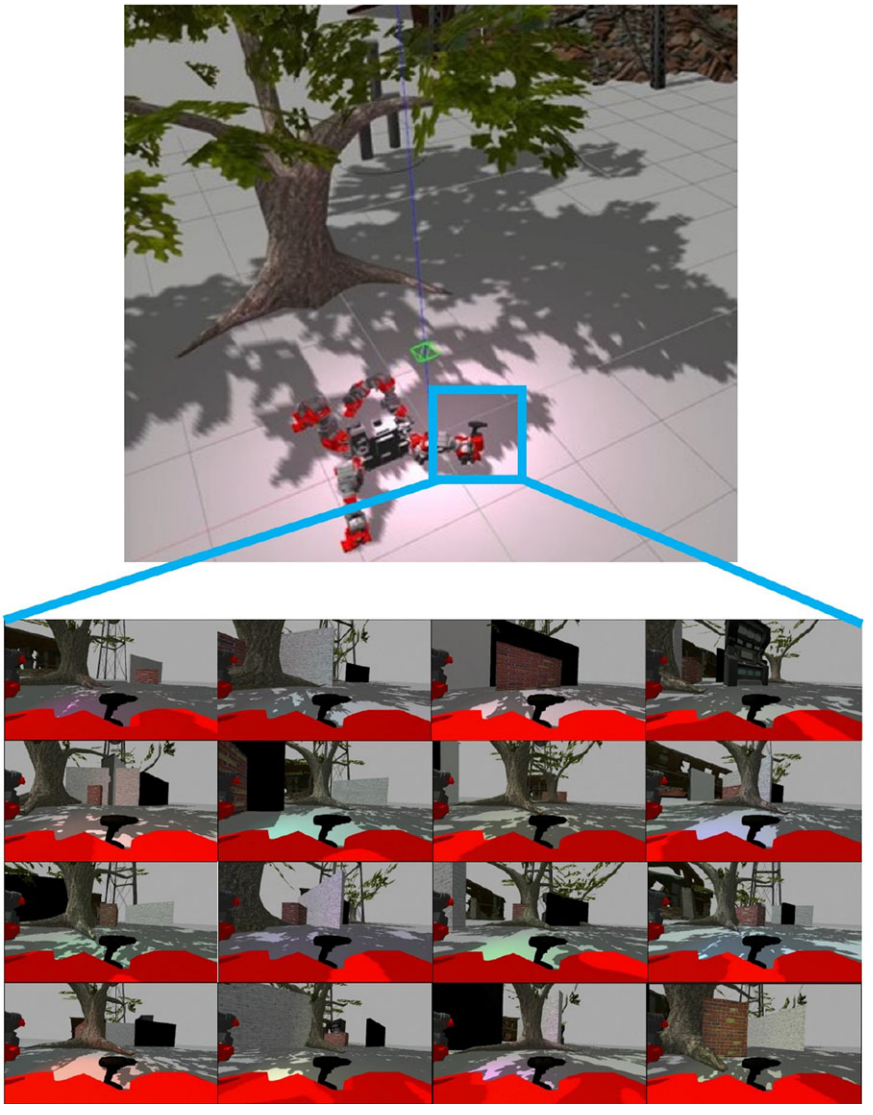


Figure 8. *Training in simulation with random light and background. The upper is overview of simulation and the lower are examples of camera view.*

system. Each subject was given five trials to grasp the target drill for both the case with and without assist system.

To begin with, success rate of reaching and grasping the target drill for all 21 subjects in both two cases is listed in Table II. Results show that the effect of assist on success rate is remarkable with the increase of 66% (from 10% to 76%) in average.

In addition to the improvement of success rate, in this paper, the index of NASA-TLX [19], which is widely known and used for the quantitative evaluation of workload of human in performing tasks is utilized as subjective scoring system for the evaluation of effect in decreasing the burden of the operators. In NASA-TLX, six main indexes of "Mental Demand" (MD), "Physical Demand" (PD), "Temporal Demand" (TD), "Own Performance" (OP), "Effort" (EF), and "Frustration" (FR) are scored to measure the workload of the operators.

https://doi.org/10.1017/S0263574721000618 Published online by Cambridge University Press

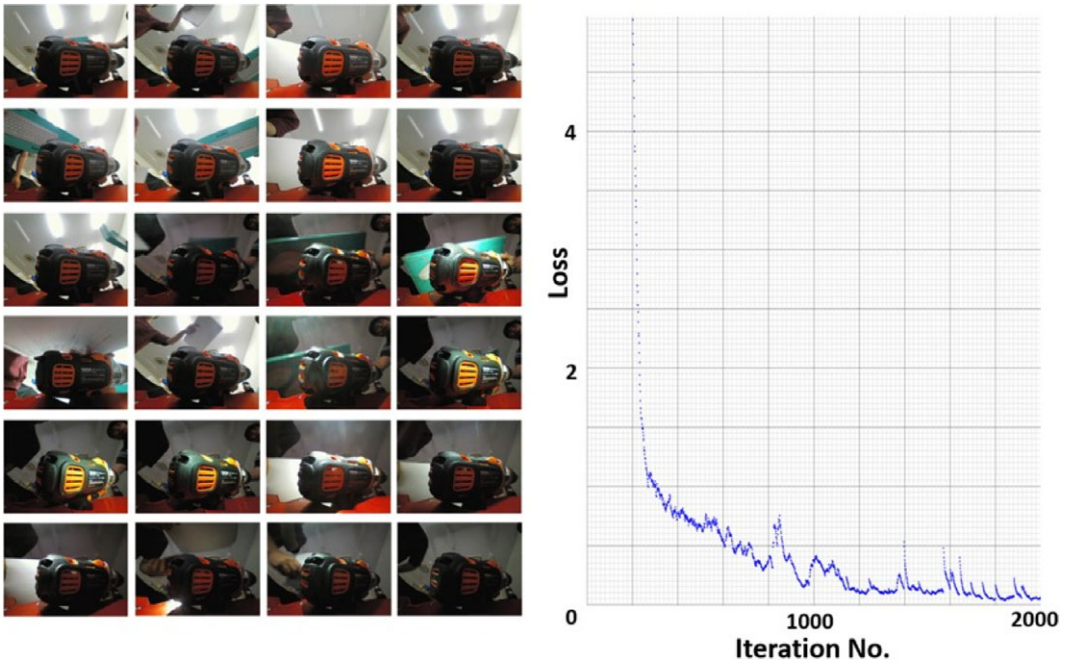


Figure 9. *Training with photos of a real electric drill (with different light and background conditions) and results.*

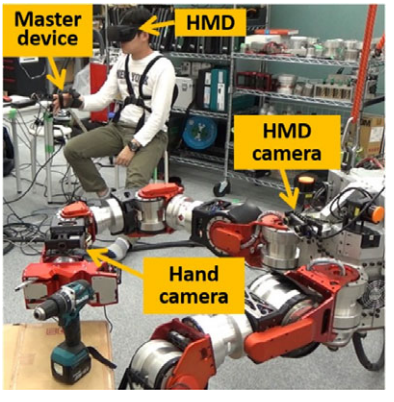


Figure 10. *Overview of the evaluation comparison experiment.*

The results of NASA-TLX for all 21 subjects in box chart and error-bar chart are shown in Figs. 11 and 12, respectively. Note that NASA-TLX is for measuring the workload of operators and lower score means less workload and better operation experience. Apparently, for all six indexes, the case with assist system has significantly fewer points, validating that it does not only contribute to the improvement of success rate in both reaching and grasping of target drills, but also decrease the workload of operators for both physical and mental aspects.

6. Conclusion and future works

This paper proposes an assist system for remote manipulation of the robot WAREC-1R in both reaching and grasping of the target object, which in this paper are electric drills with a trigger.

https://doi.org/10.1017/S0263574721000618 Published online by Cambridge University Press

Table II. *Detailed results of the experiments (N = 5 for 21 subjects).*

	Gender	Success rate (no assist) (%)	Success rate (with assist) (%)
Subject 1	Male	0	60
Subject 1	Male	0	60
Subject 2	Male	20	80
Subject 3	Male	0	80
Subject 4	Male	0	60
Subject 5	Male	0	100
Subject 6	Male	20	80
Subject 7	Female	0	80
Subject 8	Male	20	60
Subject 9	Male	0	40
Subject 10	Female	20	100
Subject 11	Female	20	80
Subject 12	Female	0	100
Subject 13	Male	0	80
Subject 14	Male	40	60
Subject 15	Female	20	60
Subject 16	Female	0	40
Subject 17	Male	20	100
Subject 18	Male	0	100
Subject 19	Male	0	60
Subject 20	Male	20	80
Subject 21	Female	0	100
Average	–	10	76

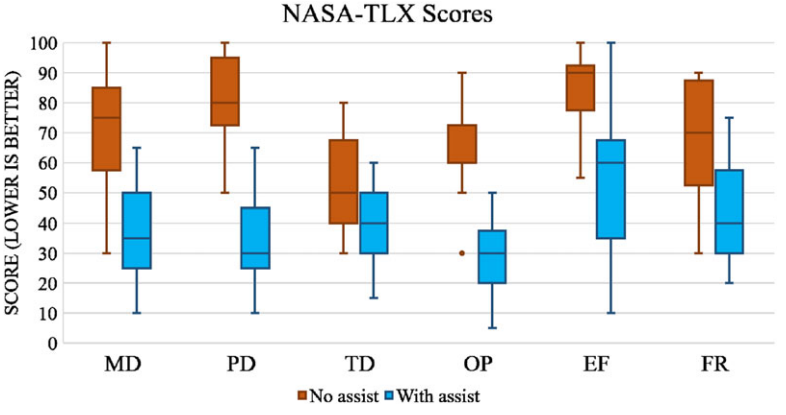


Figure 11. *NASA-TLX score in box chart. MD: Mental Demand, PD: Physical Demand, TD: Temporal Demand, OP: Own Performance, EF: Effort, FR: Frustration.*

With the introduction of hardware and devices related to manipulation as well as the whole mechatronic system configuration, in this paper, different approaches are utilized for reaching and grasping of the target electric drill. Specifically, for reaching process, Deep Q Network is applied with training in simulator instead of robots moving in real world to realize training in much higher efficiency, and object detection algorithm YOLO v3 is applied to judge the graspability in grasping process. Specifically, in

https://doi.org/10.1017/S0263574721000618 Published online by Cambridge University Press

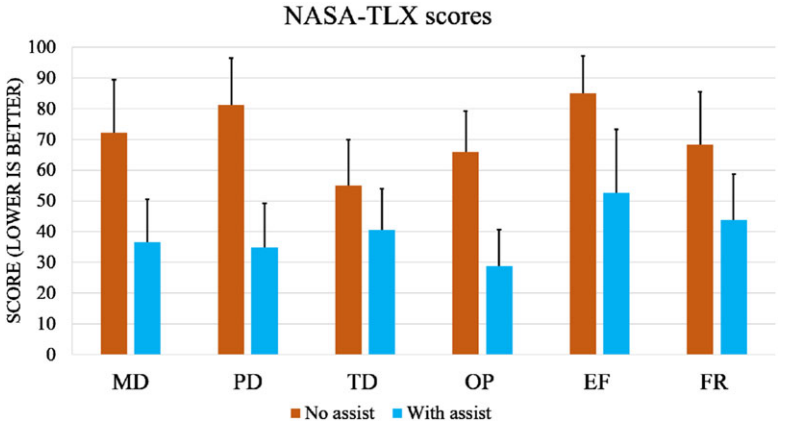


Figure 12. *NASA-TLX score in error-bar chart. MD: Mental Demand, PD: Physical Demand, TD: Temporal Demand, OP: Own Performance, EF: Effort, FR: Frustration.*

training of reaching process in simulator, the influence of light and background is considered to minimize the difference between simulation and reality, maximizing the effect of training.

Finally, results of evaluation experiments with WAREC-1R verified that the proposed assist system for remote operation improves the whole operation system for the following two points: (1) obvious increase in success rate (66% up) of grasping the target drill; (2) remarkable decrease of comprehensive workload for operators proved by comparison of NASA-TLX scores for the cases with and without the assist system.

Although assist system in this paper enables grasp of tools in complicated shape with high success rate and low workload for operators, it also has the following limits: (1) so far the training in both reaching and grasping is only available for electric drills that have a grip and a trigger; (2) similarly, the proposed assist system has only been validated to be effective for WAREC-1R and its end effector. Therefore, our prospective work would be the expansion of both target tools and robots for reaching and grasping.

Acknowledgments. This research was funded by ImPACT TRC Program of Council for Science, Technology and Innovation (Cabinet Office, Government of Japan). This study was conducted with the support of Research Institute for Science and Engineering, Waseda University; Future Robotics Organization, Waseda University, and as a part of the humanoid project at the Humanoid Robotics Institute, Waseda University. This research was also partially supported by SolidWorks Japan K. K; DYDEN Corporation; and KITO Corporation whom we thank for their financial and technical support.

Conflicts of Interest. The authors declare none.

References

[1] K. Terae, T. Matsubara, K. Hashimoto, N. Kubota, A. Namiki, X. Sun, T. Matsuzawa, A. Imai, M. Okawara, S. Kimura, K. Kumagai, K. Yamaguchi, H. Naito, K. Namura, T. Sato, M. Murakami, S. Yoshida and A. Takanishi, "Development of Disaster Response Robot for Extreme Environments (27th Report: Body Mechanism Containing Batteries and Enabling Wireless Operation with Multi-sensor System)," *The 37th Annual Conference of the Robotics Society of Japan, 1G2-01* (2019, in Japanese).

[2] Tough Robotics Challenge (TRC). https://www.jst.go.jp/impact/en/program/07.html. Accessed April 20, 2021.

[3] X. Sun, S. Hayashi, K. Hashimoto, T. Matsuzawa, Y. Yoshida, N. Sakai, A. Imai, M. Okawara, K. Kumagai, T. Matsubara, K. Yamaguchi and A. Takanishi, "Error Compensation System with Proximity Sensors for Vertical Ladder Climbing of the Robot 'WAREC-1'," *2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)* (2018) pp. 40–46. doi:10.1109/HUMANOIDS.2018.8625073.

[4] T. Matsuzawa, K. Hashimoto, X. Sun, T. Teramachi, S. Kimura, N. Sakai, Y. Yoshida, A. Imai, K. Kumagai, T. Matsubara, K. Yamaguchi, W. Tan and A. Takanishi, "Crawling Gait Generation Method for Four-Limbed Robot Based on Normalized

https://doi.org/10.1017/S0263574721000618 Published online by Cambridge University Press

Energy Stability Margin," *2017 International Symposium on Safety, Security and Rescue Robotics (SSRR), Shanghai* (2017) pp. 223–229. doi:10.1109/SSRR.2017.8088167.
[5] T. Sato, A. Imai, K. Hashimoto, A. Namiki, T. Mouri, X. Sun, T. Matsuzawa, M. Okawara, S. Kimura, K. Kumagai, T. Matsubara, K. Yamaguchi, H. Naito, K. Namura, K. Terae, M. Murakami, S. Yoshida and A. Takanishi, "Development of Disaster Response Robot for Extreme Environments (28th Report: The End-Effector of a Four-Limbed Robot Compatible with Manipulation Using Fingertips and Palm and Various Locomotion Styles)," *Proceedings of the 37th Annual Conference of the Robotics Society of Japan, 1G2-02* (2019, in Japanese).
[6] H. Chen, Y. Ji and L. Niu, "Reinforcement learning path planning algorithm based on obstacle area expansion strategy," *Intell. Serv. Rob.* **13**, 289–297 (2020). doi:10.1007/s11370-020-00313-y.
[7] D. Kim, A. Li and J. Lee, "Stable robotic grasping of multiple objects using deep neural networks," *Robotica* **39**(4), 735–748 (2020). doi:10.1017/S0263574720000703.
[8] R. Polvara, S. Sharma, J. Wan, A. Manning and R. Sutton, "Autonomous vehicular landings on the deck of an unmanned surface vehicle using deep reinforcement learning," *Robotica* **37**(11), 1867–1882 (2019). doi:10.1017/S0263574719000316.
[9] R. Yuan, F. Zhang, Y. Wang, Y. Fu and S. Wang, "A Q-learning approach based on human reasoning fornavigation in a dynamic environment," *Robotica* **37**(3), 445–468 (2019). doi:10.1017/S026357471800111X.
[10] A. Sharkawy, P. Koustoumpardis and N. Aspragathos, "Neural network design for manipulator collision detection based only on the joint position sensors," *Robotica* **38**(10), 1737–1755 (2020). doi:10.1017/S0263574719000985.
[11] A. Rajeswaran, V. Kumar, A. Gupta, G. Vezzani, J. Schulman, E. Todorov and S. Levine, "Learning complex dexterous manipulation with deep reinforcement learning and demonstrations," *Comput. Sci.* (2018). doi:10.15607/RSS.2018.XIV.049.
[12] K. Fang, Y. Zhu, A. Garg, A. Kurenkov, V. Mehta, L. Fei-Fei and S. Savarese, "Learning task-oriented grasping for tool manipulation from simulated self-supervision," *Int. J. Rob. Res.* **39**(4) (2019). doi:10.1177/0278364919872545.
[13] M. Andrychowicz, B. Baker, M. Chociej, R. Jozefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray, J. Schneider, S. Sidor, J. Tobin, P. Welinder, L. Weng and W. Zaremba, "Learning dexterous in-hand manipulation," *Int. J. Rob. Res.* **39**(1) (2018). doi:10.1177/0278364919887447.
[14] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," CoRR, abs/1804.02767 (2018).
[15] K. Hashimoto, S. Kimura, N. Sakai, S. Hamamoto, A. Koizumi, X. Sun, T. Matsuzawa, T. Teramachi, Y. Yoshida, A. Imai, K. Kumagai, T. Matsubara, K. Yamaguchi, G. Ma and A. Takanishi, "WAREC-1 - A Four-Limbed Robot Having High Locomotion Ability with Versatility in Locomotion Styles," *2017 International Symposium on Safety, Security and Rescue Robotics (SSRR)* (2017) pp. 172–178. doi:10.1109/SSRR.2017.8088159.
[16] A. Namiki, M. Gao, S. Matsushita, N. Ito, T. Tanaka, A. Ueda, Y. Murakami, S. Ikeda, T. Wada and S. Tachi, "Development of Remote Operation System of High-Power Dual Arm Robot Controlled by Telexistence FST," *Robotics and Mechatronics Conference 2012, 2P1-P04* (2012, in Japanese).
[17] A. Namiki, Y. Matsumoto, Y. Liu and T. Maruyama, "Vision-Based Predictive Assist Control on Master-Slave Systems," *2017 IEEE International Conference on Robotics and Automation (ICRA)* (2017) pp. 5357–5362. doi:10.1109/ICRA.2017.7989630.
[18] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature* **518**(7540), 529–533 (2015). doi:10.1038/nature14236.
[19] NASA TLX: TASK LOAD INDEX. https://humansystems.arc.nasa.gov/groups/TLX/. Accessed June 20, 2020.
[20] Basic theory of DQN. https://www.renom.jp/ja/notebooks/tutorial/reinforcement_learning/DQN-theory/notebook.html (in Japanese). Accessed April 20, 2021.

Appendix

Here, mechanism and algorithms of Deep Q Network [18, 20] are briefly presented as the supplementary knowledge. First of all, we define a Markov Decision Process with $S = \{s_1, s_2, ...\}$ as (finite) state space, $A = \{a_1, a_2, ...\}$ as (finite) action space, $\pi(a_t|s_t)$ as policy, $P : S \times A \times S \to [0, 1]$ as transition probability and $R : S \times A \times S \to \mathbb{R}$ as reward.

Through the interaction between the agent and environment, the goal of the agent is to find the policy π that maximizes the future rewards. In another word, for the agent the objective function is

$$\sum_{t=0}^{\infty} \gamma_t R(s_t, a_t, s_{t+1}) \tag{A1}$$

https://doi.org/10.1017/S0263574721000618 Published online by Cambridge University Press

where t is time and $\gamma \in [0, 1]$ is discount factor that determines the comparative weight between present rewards and future rewards. When there are iterations in Markov Decision Process, data obtained can be expressed by (s_r, a_r, s_{t+1}, r_t), which is called episode data and $r_t = R(s_t, a_t, s_{t+1})$. We also define

$$V^{\pi}(s) = \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t, s_{t+1}) | s_0 = s\right] \tag{A2}$$

$$Q^{\pi}(s, a) = \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t, s_{t+1}) | s_0 = s, a_0 = a\right] \tag{A3}$$

as state value function and action value function, respectively, where $\mathbb{E}_{\pi,P}$ is the average operation of the episode data when the agent takes trial and error following the policy π and transition probability P. Just as their names, state value function and action value function are the value of state s and the value of action a in the state s.

Now the objective of the reinforcement learning is to find the optimal policy π^*. Set the optimal state value function as $V^*(s) = \max_{\pi} V^{\pi}(s)$ and the optimal action value function as $Q^*(s, a) = \max_{\pi} Q^{\pi}(s, a)$, then we have $V^*(s) = \max_a Q^*(s, a)$ and

$$\begin{aligned} Q^*(s, a) &= \mathbb{E}_{s_1 \sim P(s_1|s,a)}[R(s, a, s_1) + \gamma V^*(s_1)] \\ &= \mathbb{E}_{s_1 \sim P(s_1|s,a)}[R(s, a, s_1) + \gamma \max_{a_1} Q^*(s_1, a_1)] \end{aligned} \tag{A4}$$

The optimal action value function Q^* satisfying the equation above is called Bellman equation, which is used for finding optimal policy. Regarding specific contents of action value function, taking a Finite Markov Process as instance, it can be expressed as

$$Q = \begin{pmatrix} Q(s_1, a_1) & Q(s_1, a_2) & \dots & Q(s_1, a_{|A|}) \\ Q(s_2, a_1) & Q(s_2, a_2) & \dots & Q(s_2, a_{|A|}) \\ \vdots & \vdots & \ddots & \vdots \\ Q(s_{|S|}, a_1) & Q(s_{|S|}, a_2) & \dots & Q(s_{|S|}, a_{|A|}) \end{pmatrix} \tag{A5}$$

which is called lookup tables. Apparently, when state space S and action space A get larger, the dimension of the lookup table explodes drastically, causing the famous issue of "the curse of dimensionality" with unrealsitic amount of calculation. DQN is exactly the solusion to it. Bellman equation of (A4) can be used as the objective function of DQN and the objective function of DQN can be expressed as

$$L(w) = \sum_{(s,a,s',r) \in D} (Q_w(s, a) - (R(s, a, s_1) + \gamma \max_{a1} Q_w(s_1, a_1)))^2 \tag{A6}$$

where Q_w is the Q-function that is modeled in a neural network. Set D as the episode data, we have the optimal policy by minimizing $L(w)$.

Cite this article: X. Sun, H. Naito, A. Namiki, Y. Liu, T. Matsuzawa and A. Takanishi (2022). "Assist system for remote manipulation of electric drills by the robot "WAREC-1R" using deep reinforcement learning", Robotica **40**, 365–376. https://doi.org/10.1017/S0263574721000618

https://doi.org/10.1017/S0263574721000618 Published online by Cambridge University Press