OXFORD CDE

Journal of Computational Design and Engineering, 2023, **10**, 414–424

DOI: 10.1093/jcde/qwac122
Advance access publication date: 6 December 2022
Research Article

Heave reduction of payload through crane control based on deep reinforcement learning using dual offshore cranes

Jun-Hyeok Bae[1], Ju-Hwan Cha[2,*] and Sol Ha[2]

[1]Department of Wind● Ocean Power R&D, Green Energy Institute, Mokpo-city, Jeollanam-do, Republic of Korea
[2]Department of Naval Architecture and Ocean Engineering, Mokpo National University, Muan-gun, Jeollanam-do, Republic of Korea
*Corresponding author. E-mail: jhcha@mnu.ac.kr

Abstract

Offshore operation causes the dynamic motion of offshore cranes and payload by the ocean environment. The motion of the payload lowers the safety and efficiency of the work, which may increase the working time or cause accidents. Therefore, we design a control method for the crane using artificial intelligence to minimize the heave motion of the payload. Herein, reinforcement learning (RL), which calculates actions according to states, is applied. Furthermore, the deep deterministic policy gradient (DDPG) algorithm is used because the actions need to be determined in a continuous state. In the DDPG algorithm, the state is defined as the motion of the crane and speed of the wire rope, and the action is defined as the speed of the wire rope. In addition, the reward is calculated using the motion of the payload. In this study, the heave motion of the payload was reduced by developing an agent suitable for adjusting the length of the wire rope. The heave motion of the payload was compared in between the non-learning condition of the RL-based control and proportional integral differential (PID) control; and an average payload reduction rate of 30% was observed under RL-based control. The RL-based control performed better than the PID control under learned conditions.

Keywords: deep deterministic policy gradient, reinforcement learning, ocean environment, offshore crane, crane control

List of symbols

a_t :	Action at t
a_t^* :	Optimal action at t
B :	Replay buffer (repository of states, actions, and rewards)
$C(q)$:	Constraint matrix
D_{max} :	Maximum angle
D_{target} :	Target angle
$E(G_t \mid s_t,\ a_t)$:	Probability mean of G for state and action
F :	External force matrix
G_t :	Sum of rewards up to t
$J(q)$:	Velocity transformation matrix
K_p :	Proportional coefficient of PID control
K_i :	Integral coefficient of PID control
K_d :	Differential coefficient of PID control
L :	Temporal difference error
M :	Mass and inertia matrix
N :	Three-dimensional transformation matrix
$P(s_1 \mid s_0, a_0)$:	Probability of next state for state and action
q :	Generalized coordinate matrix
$Q(s, a)$:	Q-learning to act when in a state
R_t :	Reward at time t
s_t :	State at time t
t :	Time
y_i :	Temporal difference target
γ^i :	Depreciation rate
ϵ :	Exploration noise
θ :	Weight of actor-network
λ :	Constraint force matrix
λ_L :	Learning rate
w :	Weight of critic network

1. Introduction

1.1. Background

Many operations at sea often utilize offshore cranes to handle heavy structures. During operation, offshore cranes and heavy structures move dynamically under the influence of the offshore environment such as waves, winds, and currents; this is one of the major factors hindering the safety of offshore operations. Therefore, marine operations are typically performed when the ocean environment is favorable. The crane control technique is a method that reduces the motion of offshore cranes and payloads to ensure safety. Depending on the crane control method, the dynamic motion of the payloads caused by the ocean environment can be reduced. In addition, active heave compensation (AHC) is used to reduce the motion of the payload. The AHC is a method of controlling the length of the wire rope to reduce the motion of hanging objects by analyzing the motion of the crane. However, the AHC method has a limitation; it can only be applied to specific cranes.

In this study, the control effect and possibility were analyzed by applying reinforcement learning (RL)-based control to off-

Received: June 22, 2022. **Revised:** November 7, 2022. **Accepted:** November 7, 2022

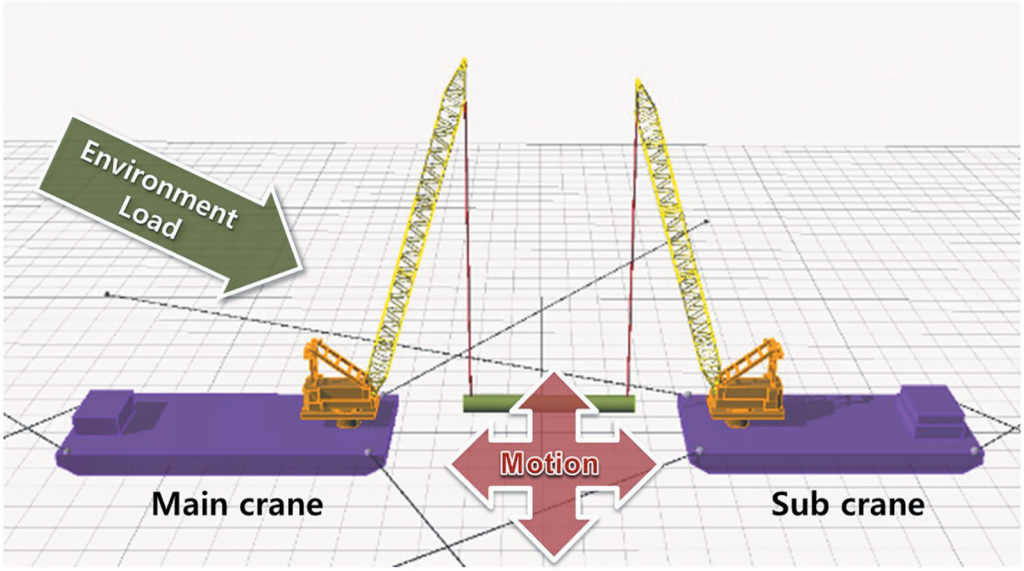

Figure 1: Operation configuration using dual cranes.

shore cranes. As shown in Fig. 1, we present a crane control method using RL to reduce the movement of the payload. The proposed crane control method is tested wherein the payload is suspended from two facing offshore cranes in an ocean environment. In addition, we performed comparisons with usually used proportional integral differential (PID) controls to evaluate control methods.

1.2. Related works

Zanjani and Mobayen (2022) studied the offshore cranes that moved containers into the sea and suggested a control method to minimize the motion of the offshore cranes; the control method was global sliding mode control (SMC). Furthermore, the crane was analyzed using kinematics to minimize motion. Li *et al.* (2020) presented a dual-loop controller of active disturbance rejection control–equivalent saturation model predictive control to solve problems such as the non-operable area of the existing AHC and correction of the non-linear characteristics of the hydraulic motor system. Liu *et al.* (2018) used fuzzy PID control as a crane control method considering the ocean environment, improved the existing fuzzy PID control algorithm, and reduced the motion of objects owing to waves using their proposed crane control. In addition, they performed a model test and compared the test results with simulation results. Chu *et al.* (2020) used a neural network algorithm to predict the movement of a ship and subsequently improved the AHC for offshore crane operations. Their algorithm was developed as a virtual prototype, and the prediction of the motion of the ship was learned and reflected in crane work in a virtual environment. Richter *et al.* (2017) present a comprehensive AHC approach proposing estimation, prediction, and control methods. The posture is estimated using the sensor and the compensation value is predicted based on the Levinson recursive least squares algorithm. Woodacre *et al.* (2015) provide the most common motion operation method with details on passive heave compensation (PHC), AHC, hybrid AHC–PHC systems, and wave synchronization systems. Zinage and Somayajula (2020) applied various controls such as proportional-derivative control, model predictive control, linear quadratic integral control, and SMC to configure effective AHC. The performance of the controllers is compared with respect to heave compensation, disturbance rejection, and noise attenuation.

Recently, control methods using RL have been studied, and RL algorithms have been developed and verified using games (Lillicrap *et al.*, 2015). In addition to the gaming sector, RL algorithms are applied and used in robotics (Gu *et al.*, 2017), natural language processing (Sharma & Kaushik, 2017), computer vision (Yun *et al.*, 2017), transportation (Farazi *et al.*, 2020), finance (Mosavi *et al.*, 2015), and healthcare; and they have a great potential in a variety of areas (Liu *et al.*, 2017). Yang *et al.* (2021) applied neural adaptive control to a 4-degrees of fredom (DOF) tower crane. Rigatos *et al.* (2022) proposed a non-linear optimal control approach for the dynamic model of the 4-DOF underactuated tower crane. Kim *et al.* (2022) proposed an adaptive predictive anti-swing control law using an empirical model after converting a data-driven model to a state space form.

Wu *et al.* (2018) applied the deep deterministic policy gradient (DDPG) algorithm to stably control the gait of a biped robot and confirmed that it effectively controlled the fall of the robot in various initial states of the biped robot. Andersson *et al.* (2021) applied RL control to increase the energy efficiency of crane operation of forestry machinery. Maamoun (2022) presented an optimal control system for actuators of offshore cranes using the Q-learning algorithm. Sierra-Garcia *et al.* (2022) applied RL to the pitch control of wind turbines. Because the energy productivity of wind power generators varies greatly with pitch control, optimal control is required. Also, considering the disadvantages of RL, pitch control was performed by combining an RL-based controller and PID regulator.

In addition, Chun *et al.* (2022) proposed a control method for determining static equilibrium and a deep RL-based method for block lifting for a crane control method that occurs during block assembly in a shipyard. Cho *et al.* (2022) applied scheduling optimization through RL to improve the productivity of block assembly lines in shipyards.

2. Crane Control Method Based on RL

Offshore cranes may cause dynamic motion owing to the ocean environment; this dynamic motion may reduce the safety of

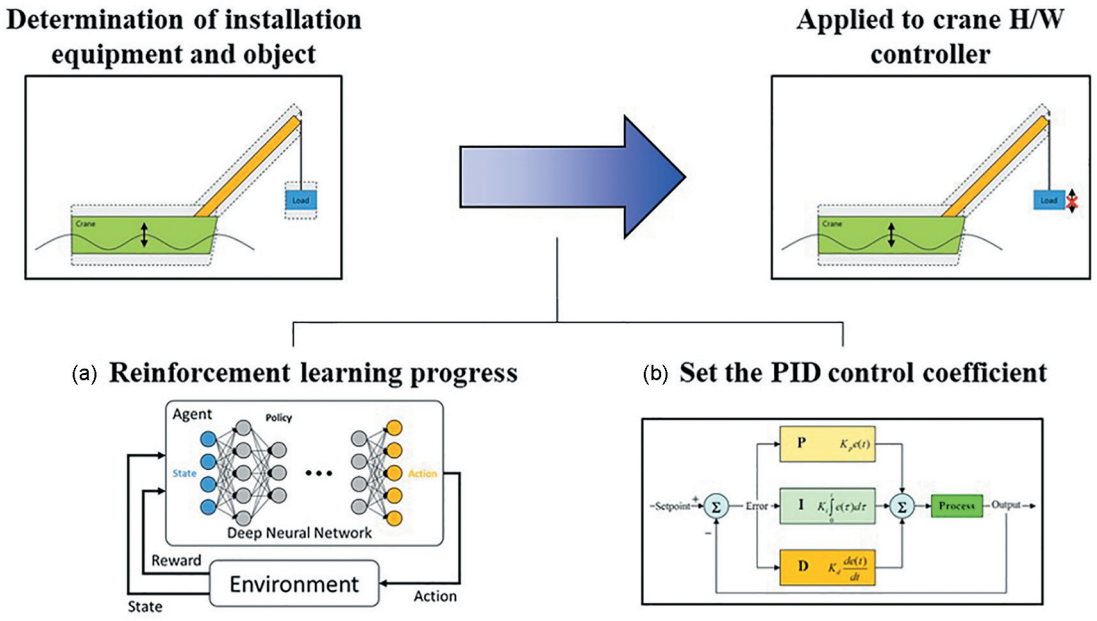

Figure 2: Comparison of control methods.

installation work or make installation difficult. Therefore, offshore operations require control to minimize the movement of the payloads in the ocean environment. Previous studies have been conducted to find solutions for specific purposes, e.g., the AHC can change the length of the wire rope to reduce the heave motion, control the ramp section to allow smooth movement, and control anti-sway to reduce sway motion.

In this study, an AHC was constructed based on RL, as shown in Fig. 2a. Control using RL can be applied to various scenarios and ocean environments by employing the learned data. However, if the installation equipment, object, and connection methods are changed, learning must be repeated.

2.1. Control method comparison

Here, Fig. 2 shows a comparison between the PID control and the control using RL. Both control methods require the determination of equipment and payloads. For PID control, we input a random coefficient and evaluate whether it is appropriate. If it is not appropriate, the optimal PID coefficient must be determined through iterative work (Johnson & Moradi, 2005). Contrastingly, in the control using RL, the state, action; and reward are defined, and when learning is performed, the algorithm determines an appropriate policy. A policy means a weight between layers in machine learning, and the policy is determined through learning.

In the case of PID control, an appropriate PID coefficient can be found through manual iterative work or optimization method. However, there is no evaluation target. Moreover, it cannot be determined whether the chosen PID coefficient is optimal; essentially, it is difficult to determine the optimal PID coefficient. Furthermore, if the operation conditions change, it is necessary to determine the corresponding optimal PID coefficient again. Artificial intelligence uses algorithms to perform tasks efficiently and therefore achieves effective results. The algorithm learns and determines an appropriate policy that matches the state, action, and reward conditions. In the future, it is expected that control using RL will be possible in various fields by expanding the scope of application of the algorithm.

2.2. DDPG algorithm

Through learning, agents in RL can interact with the environment by selecting actions and observing states and rewards without any knowledge of the kinematics. The ultimate goal of RL is to receive maximum reward; to achieve this, an action that maximizes Q must be selected, as shown in Equation (1). The expected value of the reward obtained when performing an action in the current state can be expressed by Equation (2), and the overall reward can be defined by Equation (3):

$$a_t^* = \arg\max_{a_t} Q(s_t, a_t) \tag{1}$$

$$Q(s_t, a_t) = E(G_t | s_t, a_t) \tag{2}$$

$$G_t = \sum_{i=0}^{\infty} \gamma^i R_{t+i}\,. \tag{3}$$

Table 1 lists the DDPG algorithm (Gao *et al.*, 2020; Lillicrap *et al.*, 2015; Zinage & Somayajula, 2021). The algorithm initializes the actor, critic, target actor, target critic, and replay buffer. Next, it receives the state from the environment and accordingly determines the policy action. Subsequently, noise (Uhlenbeck and Ornstein, 1930) is added to the action to obtain a better reward. Thus, it performs actions in the environment, receives rewards in the next state, and stores them in the replay buffer. When a certain amount of data is accumulated in the replay buffer, the data are extracted according to the batch size and learning is performed by the mean square error method. And the actor and critic networks learn and update each target network.

3. RL Configuration

3.1. Dynamic analysis conditions

To compare the control methods, pipe installation operation using 160 and 150 ton crane barge was selected. The offshore crane was modeled as a multibody composed of several rigid bodies and joints as shown in Fig. 3, and hydrodynamic force (Seo *et al.*, 2018) and mooring force (Bae *et al.*, 2021) were considered as external forces. Multibody system equations of motion should be applied to

Table 1: DDPG algorithm.

1: Randomly initialize actor (a) and critic (Q) with weights θ and w
2: Initialize target network a^- and Q^- with weights θ^- and w^-
3: Initialize replay buffer B
4: **for** episode $= 1, M$ **do**
5: Initialize a random process ϵ for action exploration
6: Receive initial observation state s_1
7: **for** $t = 1, T$ **do**
8: Select action $a_{\theta,t} + \epsilon_t$ (ϵ_t is exploration noise)
9: Execute action at and observe reward R_t and new state s_{t+1}
10: Store transition $(s_t,\ a_t,\ R_t,\ s_{t+1})$ in B
11: Sample a random minibatch of N transitions $(s_i,\ a_i,\ R_i,\ s_{i+1})$ from B
12: Set $y_i = R_i + \gamma Q_{w^-}(s_{i+1}, a_{\theta^-,i+1})$
13: Update critic by minimizing the loss:
$L = \frac{1}{N}\sum_{i=1}^{N}(Q_w(s_i, a_{\theta,t}) - y_i)^2$
14: Update the actor policy using the sampled policy gradient:
$\nabla_\theta J_\theta \cong \sum_{t=0}^{\infty} \int_{s_0,s_1,s_2,\cdots,s_t} \nabla_\theta a_{\theta,t} \nabla_{a_{\theta,t}} Q(s_t, a_{\theta,t}) P(s_0)P(s_1|s_0, a_{\theta,0})P(s_2|s_1, a_{\theta,1})\cdots P(s_t|s_{t-1}, a_{\theta,t-1}) ds_0, s_1, s_2, \cdots, s_t$
15: Update the target networks:
$\theta^- \leftarrow (1-\lambda_L)\theta^- + \lambda_L\theta\ w^- \leftarrow (1-\lambda_L)w^- + \lambda_L w$
16: **end for**
17: **end for**

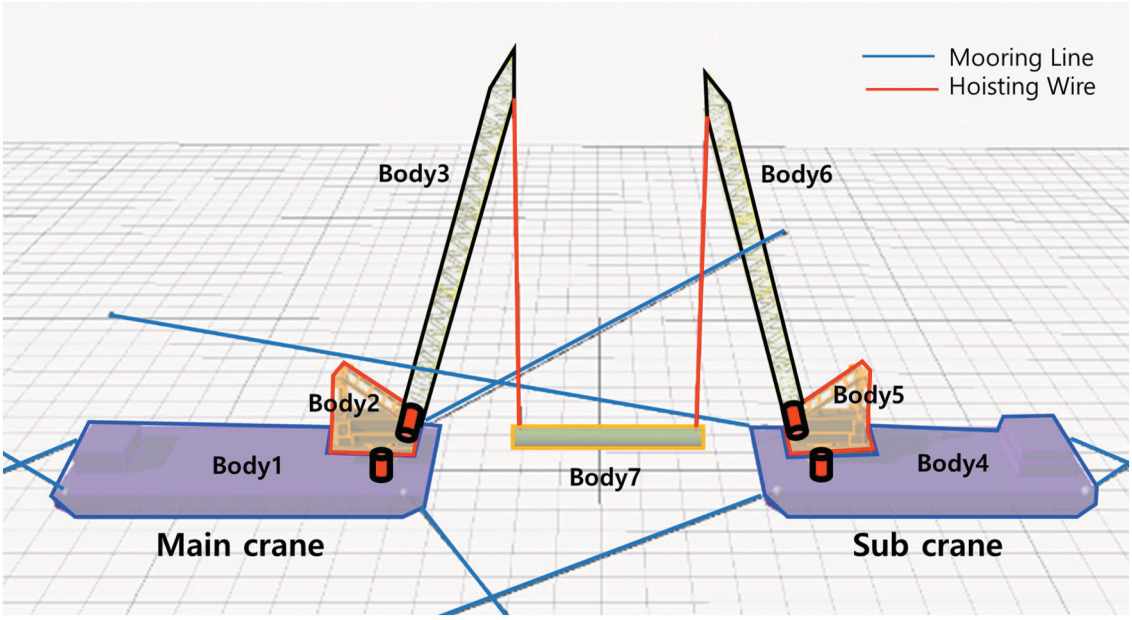

Figure 3: Construction of multibody system for dual crane operation.

objects that are connected to multiple rigid bodies with joints and wire ropes, and analysis was performed using the augmented formulation as in Equation (4) among many multibody system equations of motion (Cha *et al.*, 2012; Shabana, 1994; Shabana, 2005).

$$\begin{cases} \bar{\mathbf{M}}(\mathbf{q})\ddot{\mathbf{q}} + \mathbf{C}_\mathbf{q}^{\mathrm{T}}(\mathbf{q})\lambda = \bar{\mathbf{F}}(\mathbf{q}, \dot{\mathbf{q}}) - \bar{\mathbf{k}}(\mathbf{q}, \dot{\mathbf{q}}) \\ \mathbf{C}(\mathbf{q}) = \mathbf{0} \end{cases} \tag{4}$$

$$\bar{\mathbf{M}}(\mathbf{q}) = \mathbf{J}(\mathbf{q})^{\mathrm{T}}\mathbf{M}\mathbf{J}(\mathbf{q})$$

$$\bar{\mathbf{F}}(\mathbf{q}, \dot{\mathbf{q}}) = \mathbf{J}(\mathbf{q})^{\mathrm{T}}\mathbf{F}(\mathbf{q}, \dot{\mathbf{q}})$$

$$\bar{\mathbf{k}}(\mathbf{q}, \dot{\mathbf{q}}) = \mathbf{J}(\mathbf{q})^{\mathrm{T}}\mathbf{M}\dot{\mathbf{J}}(\mathbf{q})\dot{\mathbf{q}} + \mathbf{J}(\mathbf{q})^{\mathrm{T}}\mathbf{N}(\dot{\mathbf{q}})\mathbf{J}(\mathbf{q})\dot{\mathbf{q}}$$

3.2. Analysis system configuration

Simulation and RL were developed in different languages, the data were exchanged between the two languages using data communication, and the data input/output were defined as shown in Fig. 4.

When initialization is in progress in the simulation system, the analysis initial information is sent to the RL, and the RL creates an agent based on the analysis initial information. At the same time, the simulation system calculates the static analysis and sends the calculated value back to the RL. When RL calculates an action that matches the policy and passes it to the simulation, the simulation performs an analysis that reflects the control. The simulation result (6 degree of freedom) is sent to the RL, and the RL performs learning with state, action, reward, and new state. When the simulation is finished during iteration, the analysis system initializes

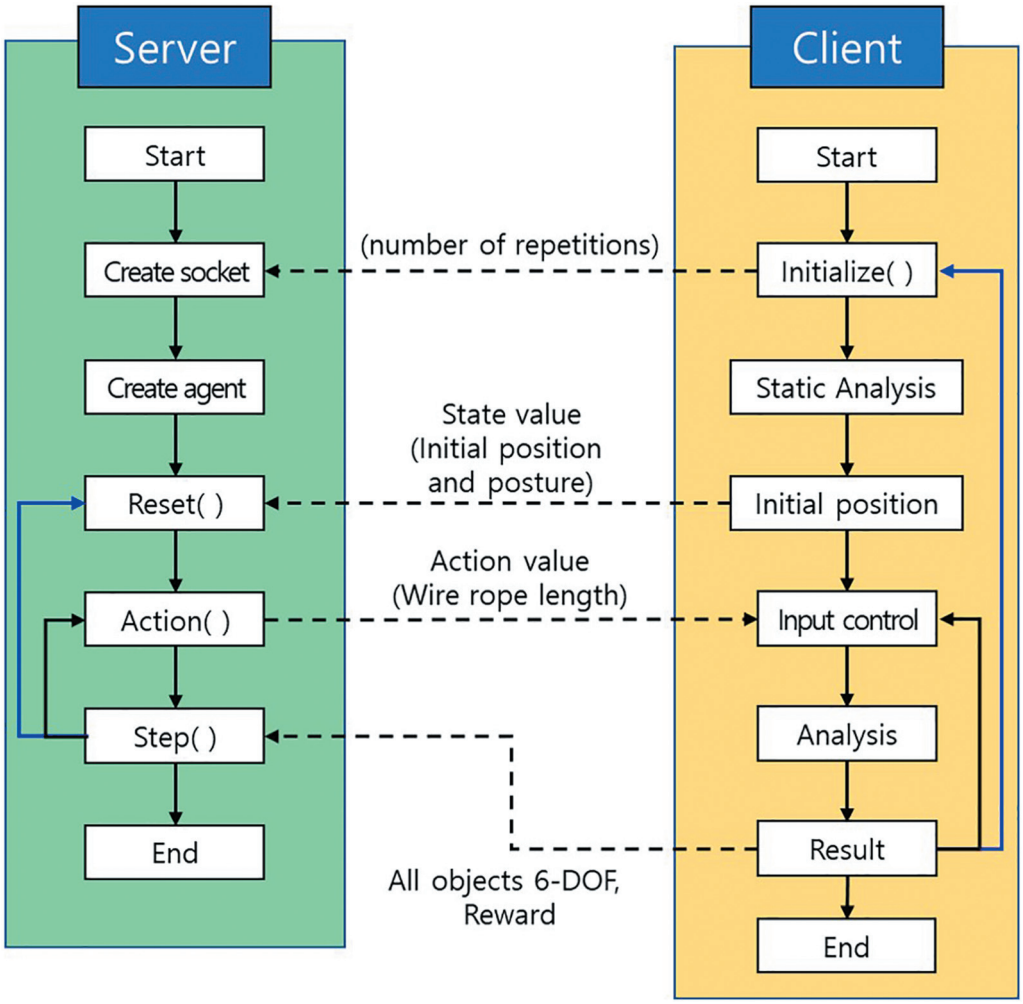

Figure 4: Configuration of data flow.

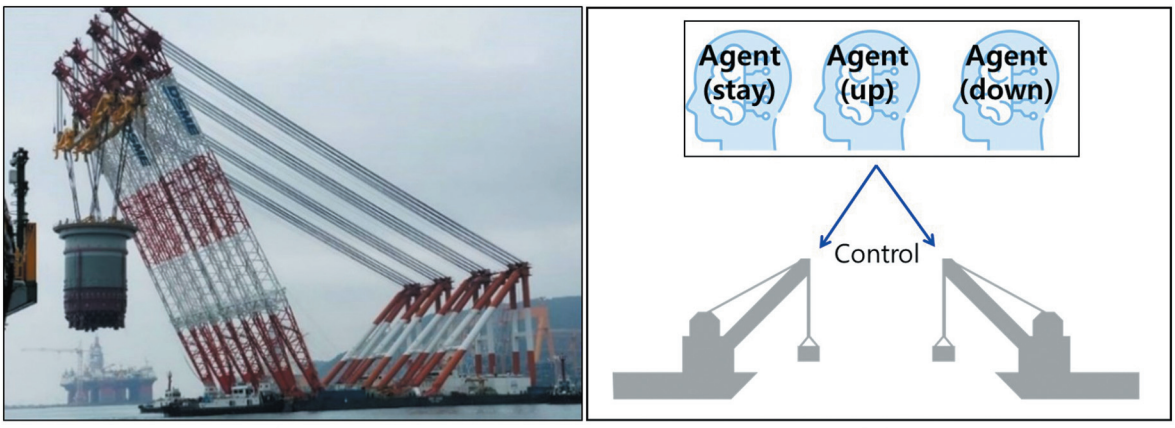

Figure 5: Configuration of parallel crane and agent.

and performs re-analysis, and the RL proceeds with learning according to the analysis system.

3.3. Definition of state, action, and reward

Defining precise goals and rewards improves learning effectiveness. As for the learning condition, the cases of lifting, stopping, and lowering the payload were divided, and learning was performed individually. There are 18 items for status and two items for action, as follows:

(i) State: (payload) heave, heave velocity, pitch, and pitch velocity; (main crane) roll, roll velocity, pitch, pitch velocity, yaw, and yaw velocity; (sub-crane) roll, roll velocity, pitch, pitch

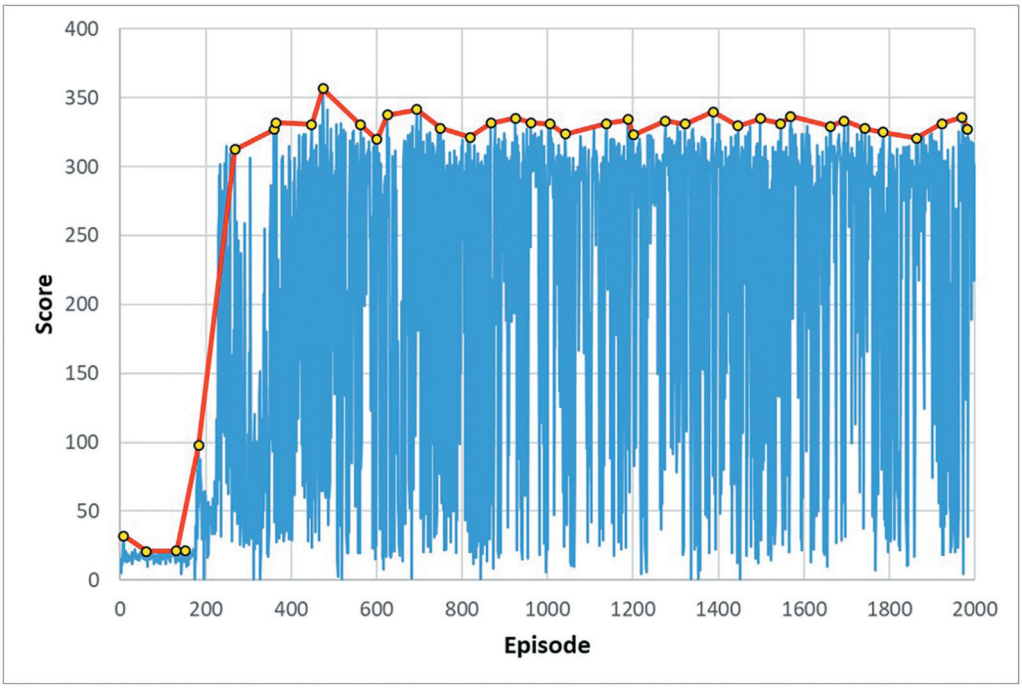

Figure 6: Change of score according to episode.

velocity, yaw, and yaw velocity; (main crane) speed of wire rope; and (sub-crane) speed of wire rope.

(ii) Action: (main crane) speed of wire rope and (sub-crane) speed of wire rope.

The configuration of the main crane, sub-crane, and payload is shown in Fig. 1. The state and action consist of data and minimum items that can be acquired for the application to actual cranes. Policy of RL determines the action of receiving a high reward; therefore, it is important to define the reward function. The reward is a value between 0 and 1, and the difference between the target value and current value is calculated using Equation (5). Since the reward consists of several sums, it is difficult to know the exact previous state with the motion of the crane and the object, so an action is added to the state.

$$R = \frac{e^{-|x-D_{\text{target}}|} - e^{-D_{\max}}}{1 - e^{-D_{\max}}}, \tag{5}$$

Where x is the current motion, D_{target} is the target motion, and $D_{\max}$ is the maximum motion. When the maximum value is large, it tends to fail to converge, and when the maximum value is small, it may be difficult to proceed with learning. Therefore, in this study, the maximum motion was defined as 1.2 times the amplitude of the motion or velocity.

The reward function is composed of heave motion, pitch motion, and heave velocity. The reward function is composed of heave motion, pitch motion, and heave velocity. The total reward is calculated as the sum of the ratio of heave motion (50%), pitch motion (20%), and heave velocity (30%), as shown in Equation (6):

$$R_{\text{total}} = 0.5\, R_{\text{heave}} + 0.2\, R_{\text{pitch}} + 0.3\, R_{\text{heave velocity}}. \tag{6}$$

3.4. Learning conditions

Figure 5a shows a case where a crane is connected to perform one operation. In this case, one controller should control two cranes at the same time. If the reward is specified, learning is difficult; however, it can be applied to various environments. Conversely, if the reward is set comprehensively, learning is easy but difficult to apply to various environments. As shown in Fig. 5b, one agent trained with one target, so three agents are required for stopping, winding, and unwinding.

To adapt the learned data to various ocean environments, three wave loads were superimposed. The wave load used was an irregular wave; its spectrum was Pierson Moskowitz spectrum, wave height was 1 m, period was 4 s, and directions were 0, 45, and 90°. In the initial state, the payload was suspended in a position wherein the two crane barges faced each other. A wave load was created for 0–20 s and training was performed to reduce the heave motion for 20–60 s.

Subsequently, RL was learned by repeating 1000 times; the numerical analysis time interval was 0.01 s, control time interval was 0.1 s, and numerical analysis time was 60 s. Because the algorithm adds noise to the behavior during training, the reward is not constant; and the policy with the maximum reward is used. Therefore, as shown in Fig. 6, the maximum reward converges after about 500 times; however, 1000 times were performed to determine a better condition. In addition, the analysis time interval was set to 0.01 s using the numerical analysis method, and the control time interval was set to 0.1 s to apply to the actual crane.

4. Comparison of the Results of the Two Controls

To compare the control method, this study was compared with the commonly used PID control. The coefficient was selected as a value with a large decrease in motion through several iterations. As a result, the PID coefficients were set to 0.125 for K_p, 0.0001 for K_i, and 0.0045 for K_d. Both controls were set under the same conditions and tested in various environments.

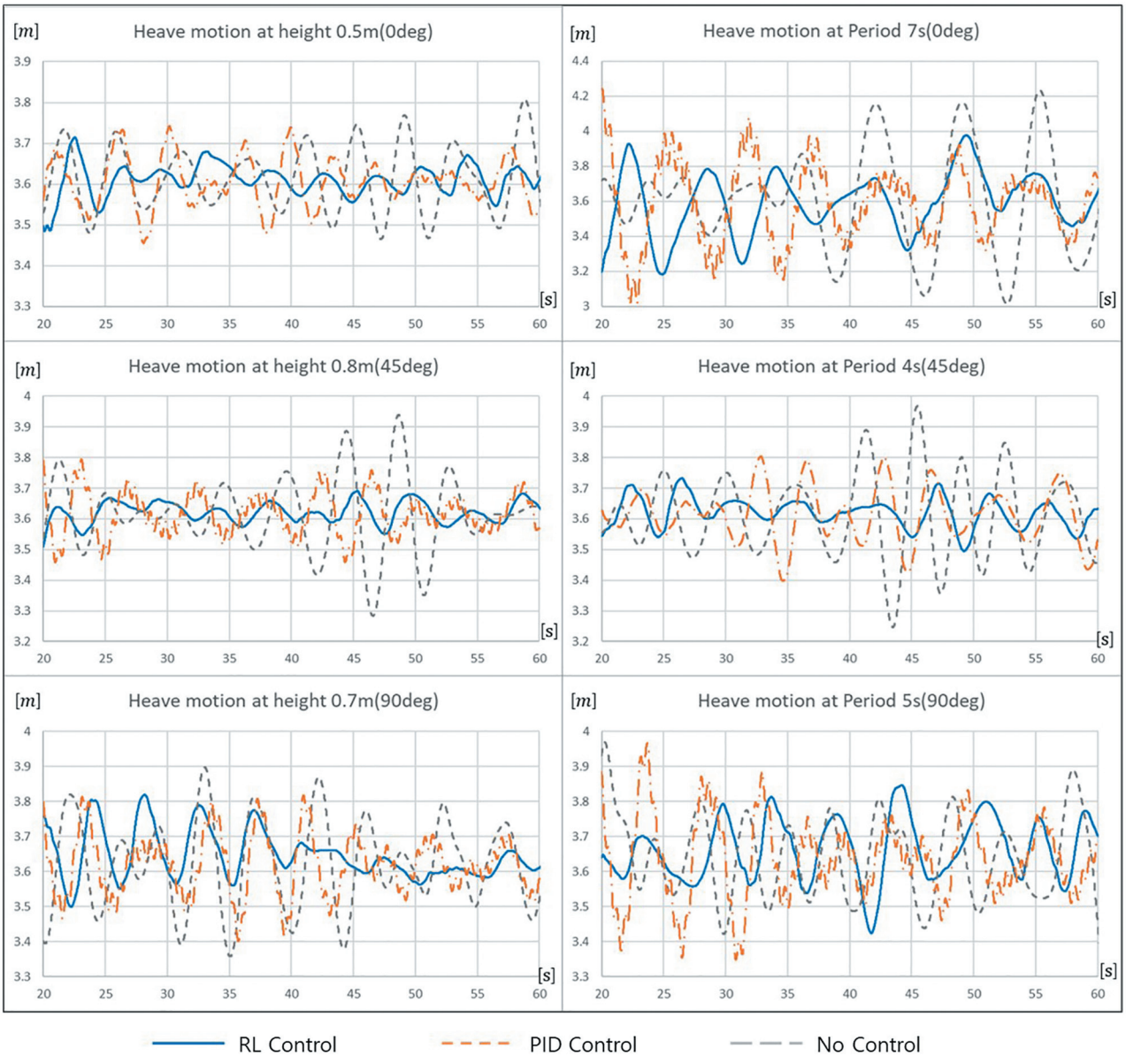

Figure 7: Heave motion by period and wave height.

4.1. Control to hold the position

The load conditions were classified by wave direction, wave height, and wave period, as follows. The wave directions were 0, 45, and 90°; the wave heights were 0.1–1.2 m with 0.1 m intervals; and the periods were 2–10 s with 1 s intervals. The learned control and PID control were compared. The PID control coefficients were found to be 0.125, 0.0001, and 0.0045 for K_p, K_i, and K_d, respectively. In addition, the results of the two control groups were compared by calculating their corresponding standard deviation.

Here, Fig. 7 shows a graph of the payload compared by changing the wave direction, wave height, and period. In the graph, the blue line represents the learned control, the orange line represents the PID control, and the gray line represents no control. The standard deviation of the heave motion was calculated for each graph, and in the case of wave height of 0.5 m and wave direction 0 deg, RL was 0.03616, PID was 0.06043, and no-control was 0.09909, and RL control decreased the heave motion the most. The standard deviation comparison is shown in Fig. 8.

Here, Fig. 8a shows the standard deviation for each direction and wave height for a period of 4 s. Figure 8b shows the standard deviation for each direction and period when the wave height was 1 m. The x-axis denotes the wave height or period and the y-axis denotes the standard deviation. We confirmed that the heave motion tended to decrease more significantly in the learned control than in the PID control, and the effect of the PID control was better than that of the learned control over a long period. These trends are observed because the PID control is effective for a long period, whereas the learned control is effective for a short period in the learned condition.

In the case of the standard deviation according to wave height, the maximum reduction rates were 55.8, 65.7, and 47.1% at 0, 45, and 90° in the direction of the ocean environment, and the

Figure 8: Standard deviation of heave by wave height.

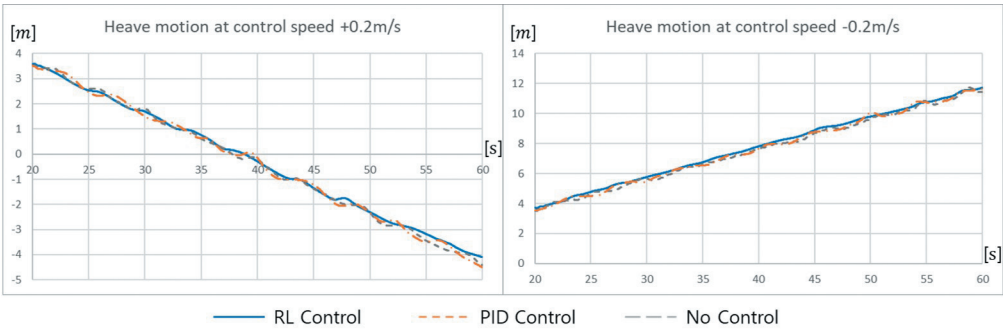

Figure 9: Heave motion by wire rope speed.

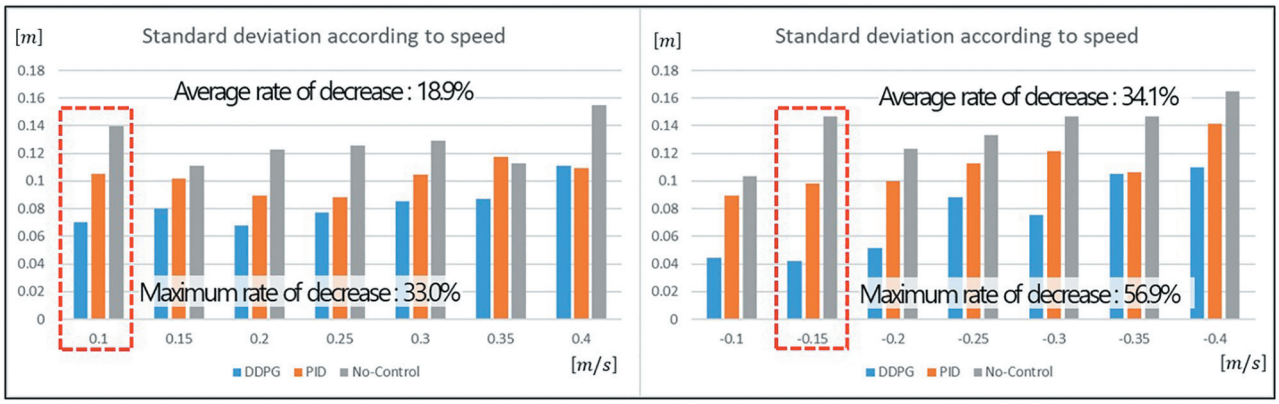

Figure 10: Standard deviation of heave by wire rope speed.

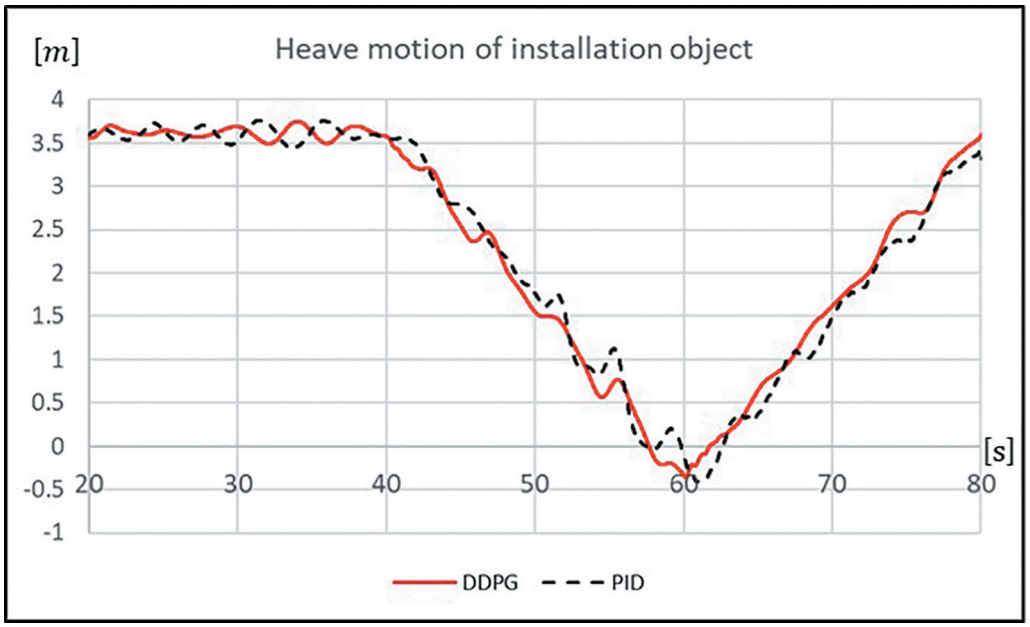

Figure 11: Heave motion of the payload.

average rate of decrease was 35.0, 37.5, and 17.8%, respectively. In the case of the standard deviation according to the period, the maximum rate of decrease was 75.5, 64.8, and 62.2 at 0, 45, and 90° in the direction of the ocean environment, and the average rate of decrease was 16.2, 0.3, and 0.3%, respectively.

4.2. Control of wire rope speed

The learned and PID controls were compared by changing the control speed. The control speed was −0.2 m/s for winding and 0.2 m/s for unwinding. The control speed for performing the evaluation was 0.1–0.4 m/s with 0.05 m/s intervals for winding and −0.1 to −0.4 m/s with 0.05 m/s intervals for unwinding.

The learned control and PID control were compared using the standard deviation. However, because the speeds were different, it was not possible to compare the two using standard deviation. Therefore, a trend line was obtained from the results and the standard deviation was calculated as the difference from the trend line.

Here, Fig. 9 shows the heave motion of the payload and the winding and unwinding of the wire rope. The blue line represents the learned control, the orange line represents the PID control, and the gray line represents no control. The heave motion of the payload is almost linear and compared to the standard deviation of the distance from the trend line. For winding, the standard deviation is 0.0681 for RL control, 0.0893 for PID control, and 0.1227 for no-control. For unwinding, the standard deviation is 0.0681 for RL control, 0.0893 for PID control, and 0.1227 for no-control. Therefore, RL control reduced 23.7% of winding and 48.7% of unwinding compared to PID control.

Here, Fig. 10 shows the standard deviation according to the speed in the winding and unwinding states. It can be observed that most of the standard deviation values tended to decrease. In the case of unwinding the wire rope, the maximum and average rates of decrease were 33.0% at 0.1 m/s and 18.9%, respectively. In the case of the winding wire rope, the maximum and average rates of decrease were 56.9% at −0.15 m/s and 34.1%, respectively.

4.3. Compare by scenario

Learning about the state of winding and unwinding of the wire ropes and the state of stopping was applied to the scenario. This

scenario was defined as creating a wave load for 0–20 s, stopping for 20–40 s, unwinding the wire rope for 40–60 s, and winding the wire rope for 60–80 s. The ocean environment was the same as that used for the learning.

Here, Fig. 11 shows the heave motion of the payload. The tendency of the heave motion to decrease much more significantly in the learned control than in PID control was confirmed by applying the agent learned from three different courses to one simulation. The rates of decrease were 26.7, 30.8, and 36.3% in the periods of 20–40, 40–60, and 60–80 s.

5. Conclusions

In this study, control was performed by applying deep RL to the installation work employing two offshore cranes. The DDPG algorithm was used to select the optimal action according to the state. The state was defined as the motion of the payload, motion of the offshore crane, and wire rope speed; the action was defined as the wire rope speed of the offshore crane. Learning was conducted separately according to specific circumstances. Verification was performed by changing the ocean environment. In the ocean environment, the wave height, wave direction, and wave period were changed and compared with the results of the PID control.

When the payload was stopped, the standard deviations according to the wave height in the 0, 45, and 90° directions were 55.8, 65.7, and 47.1%, respectively, and the standard deviations according to the period were 75.5, 64.8, and 62.2%, respectively. In the case of unwinding the wire rope, the maximum rate of decrease was 33.0% at 0.1 m/s. For the wire rope, the maximum rate of decrease in tension was 56.9% at −0.15 m/s. When applied to this scenario, the reduction rate decreased to 26.7, 30.8, and 36.3% in 20–40, 40–60, and 60–80 s.

Consequently, the tendency of the heave motion to decrease much more significantly in the learned control than in PID control was confirmed. The RL-based control was effective in the learning condition, and the motion of the payload was minimized by learning the environment of the installation area. In the future, we plan to conduct a study to determine a learning condition with good control effect for a longer cycle and control an actual crane by applying RL.

Acknowledgments

This work was supported by the New & Renewable Energy of the Korea Institute of Energy Technology Evaluation and Planning (KETEP) grant funded by the Korea Government Ministry of Knowledge Economy (No. 20213030020280) and Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea Government Ministry of Trade, Industry and Energy (MOTIE) (p0001968, The Competency Development Program for Industry Specialist).

Conflict of interest statement

None declared.

References

Andersson, J., Bodin, K., Lindmark, D., Servin, M., & Wallin, E. (2021). Reinforcement learning control of a forestry crane manipulator. In *Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems*(pp. 2121–2126). IROS. https://doi.org/10.1109/IROS51168.2021.9636219.

Bae, J., Cha, J., Seo, M. G., Lee, K., Lee, J., & Ku, N. (2021). Experimental study on development of mooring simulator for multi floating cranes. *Journal of Marine Science and Engineering*, **9**(3), 344. https://doi.org/10.3390/jmse9030344.

Cha, J. H., Park, K. P., & Lee, K. Y. (2012). Development of a simulation framework and applications to new production processes in shipyards. *Computer-Aided Design*, **44**(3), 241–252. https://doi.org/10.1016/j.cad.2011.06.010.

Cho, Y. I., Nam, S. H., Cho, K. Y., Yoon, H. C., & Woo, J. H. (2022). Minimize makespan of permutation flowshop using pointer network. *Journal of Computational Design and Engineering*, **9**(1), 51–67. https://doi.org/10.1093/jcde/qwab068.

Chu, Y., Li, G., & Zhang, H. (2020). Incorporation of ship motion prediction into active heave compensation for offshore crane operation. In *Proceedings of the 2020 15th IEEE Conference on Industrial Electronics and Applications*(pp. 1444–1449). ICIEA. https://doi.org/10.1109/ICIEA48937.2020.9248283.

Chun, D. H., Roh, M. I., & Lee, H. W. (2022). Automation of crane control for block lifting based on deep reinforcement learning. *Journal of Computational Design and Engineering*, **9**(4), 1430–1448. https://doi.org/10.1093/jcde/qwac063.

Farazi, N. P., Ahamed, T., Barua, L., & Zou, B. (2020). Deep reinforcement learning and transportation research: A comprehensive review. arXiv:2010. 06187. https://doi.org/10.48550/arXiv.2010.06187.

Gao, J., Ye, W., Guo, J., & Li, Z. (2020). Deep reinforcement learning for indoor mobile robot path planning. *Sensors*, **20**(19), 5493. https://doi.org/10.3390/s20195493.

Gu, S., Holly, E., Lillicrap, T., & Levine, S. (2017). Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *Proceedings of the IEEE international conference on robotics and automation*(pp. 3389–3396). ICRA. https://doi.org/10.1109/ICRA.2017.7989385.

Johnson, M. A., & Moradi, M. H. (2005). *PID control*. Springer.

Kim, G. H., Yoon, M., Jeon, J. Y., & Hong, K. S. (2022). Data-driven modeling and adaptive predictive anti-swing control of overhead cranes. *International Journal of Control, Automation and Systems*, **20**(8), 2712–2723. https://doi.org/10.1007/s12555-022-0025-8.

Li, Z, Ma, X, Li, Y, Meng, Q, & Li, J (2020). ADRC-ESMPC active heave compensation control strategy for offshore cranes. *Ships and Offshore Structures*, **15**(10), 1098–1106. https://doi.org/10.1080/17445302.2019.1703388.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tasaa, Y., Silver, D., & Wierstra, D. (2015). Continuous control with deep reinforcement learning. *arXiv preprint arXiv:*1509.02971. https://doi.org/10.48550/arXiv.1509.02971.

Liu, X., Li, W., Wang, W., & Xu, Z. (2018). Control for the new harsh sea conditions salvage crane based on modified fuzzy PID. *Asian Journal of Control*, **20**(4), 1582–1594. https://doi.org/10.1002/asjc.1707.

Liu, Y., Logan, B., Liu, N., Xu, Z., Tang, J., & Wang, Y. (2017). Deep reinforcement learning for dynamic treatment regimes on medical registry data. In *Proceedings of the IEEE International Conference on Healthcare Informatics*(pp. 380–385). ICHI. https://doi.org/10.1109/ICHI.2017.45.

Maamoun, K. S. A. (2022). Impact control for offshore crane in load-landing operations using reinforcement learning. Master's Thesis, ING - Scuola di Ingegneria Industriale e dell'Informazione.

Mosavi, A., Ghamisi, P., Faghan, Y., & Duan, P. (2020). Comprehensive review of deep reinforcement learning methods and applications in economics. *Mathematics*, **8**(10), 1640. https://doi.org/10.3390/math8101640.

Richter, M., Schaut, S., Walser, D., Schneider, K., & Sawodny, O. (2017). Experimental validation of an active heave compensation system: Estimation, prediction and control. *Control Engineering Practice*, **66**, 1–12. https://doi.org/10.1016/j.conengprac.2017.06.005.

Rigatos, G., Abbaszadeh, M., & Pomares, J. (2022). Nonlinear optimal control for the 4-DOF underactuated robotic tower crane. *Autonomous Intelligent Systems*, **2**(1), 1–30. https://doi.org/10.1007/s43684-022-00040-4.

Seo, M. G., Kim, N., Ha, Y. J., Park, I. B., Nam, B. W., Lee, K., & Sung, H. G. (2018). Experimental and numerical analysis of installation process using dual floating crane vessel. In *Proceedings of the Thirteenth ISOPE Pacific/Asia Offshore Mechanics Symposium*.

Shabana, A. A. (1994). *Computational dynamics*. John Wiley & Sons.

Shabana, A. A. (2005). *Dynamics of multibody systems*. (3rd ed.). Cambridge University Press.

Sharma, A. R., & Kaushik, P. (2017). Literature survey of statistical, deep and reinforce-ment learning in natural language processing. In *Proceedings of the International Conference on Computing, Communication and Automation*(pp. 350–354). ICCCA. https://doi.org/10.1109/CCAA.2017.8229841.

Sierra-Garcia, J. E., Santos, M., & Pandit, R. (2022). Wind turbine pitch reinforcement learning control improved by PID regulator and learning observer. *Engineering Applications of Artificial Intelligence*, **111**, 104769. https://doi.org/10.1016/j.engappai.2022.104769.

Uhlenbeck, G. E., & Ornstein, L. S. (1930). On the theory of the Brownian motion. *Physical Review*, **36**(5), 823. https://doi.org/10.1103/PhysRev.36.823.

Woodacre, J. K., Bauer, R. J., & Irani, R. A. (2015). A review of vertical motion heave compensation systems. *Ocean Engineering*, **104**, 140–154. https://doi.org/10.1016/j.oceaneng.2015.05.004.

Wu, X., Liu, S., Zhang, T., Yang, L., Li, Y., & Wang, T. (2018). Motion control for biped robot via DDPG-based deep reinforcement learning. In *Proceedings of the 2018 WRC Symposium on Advanced Robotics and Automation*(pp. 40–45). WRC SARA. https://doi.org/10.1109/WRC-SARA.2018.8584227.

Yang, T., Sun, N., & Fang, Y. (2021). Neuroadaptive control for complicated underactuated systems with simultaneous output and velocity constraints exerted on both actuated and unactuated states. *IEEE Transactions on Neural Networks and Learning Systems*. https://doi.org/10.1109/TNNLS.2021.3115960.

Yun, S., Choi, J., Yoo, Y., Yun, K., & Young, C. J. (2017). Action-decision networks for visual tracking with deep reinforcement learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*(pp. 2711–2720).

Zanjani, M. S., & Mobayen, S. (2022). Anti-sway control of offshore crane on the surface vessel using global sliding mode control. *International Journal of Control*, **95**, 2267–2278. https://doi.org/10.1080/00207179.2021.1906447.

Zinage, S., & Somayajula, A. (2020). A comparative study of different active heave compensation approaches. *Ocean Systems Engineering*, **10**(4), 373. http://dx.doi.org/10.12989/ose.2020.10.4.373.

Zinage, S., & Somayajula, A. (2021). Deep reinforcement learning based controller for active heave compensation. *IFAC-PapersOnLine*, **54**(16), 161–167. https://doi.org/10.1016/j.ifacol.2021.10.088.