Viewpoint Selection without Subject Experiments for Teleoperation of Robot Arm in Reaching Task Using Reinforcement Learning

Haoxiang Liu[1], Ren Komatsu[1], Hanwool Woo[1], Yusuke Tamura[2], Atsushi Yamashita[1], and Hajime Asama[1]

***Abstract*— In this study, we proposed a method to evaluate the viewpoint of a robot arm in a reaching movement using reinforcement learning. The optimal viewpoint for operators in teleoperation was studied by conducting a subject experiment. However, in some special situations, such as inside the pedestal of a nuclear plant crushed in a disaster, the lack of environmental information makes it challenging to prepare the subject experiment in advance. In addition, individual differences cannot be eliminated by conducting the subject experiment. In this study, we used reinforcement learning to select viewpoints and found that the world model inspired by the prediction function of the brain exhibited similar performance to that of humans in the reaching motion of a robot arm. This study demonstrated that the world model can evaluate viewpoints using reinforcement learning in the reaching task.**

I. INTRODUCTION

After the Fukushima Daiichi Nuclear Power Plant accident of the Tokyo Electric Power Company, Incorporated (TEPCO), which occurred in March 2011, fuel debris was left inside the pedestal of the plants. Because humans could not enter directly owing to high-level radiation, it became challenging to investigate, operate internal equipment, and remove the fuel debris. Currently, efforts are underway to decommission the plant. One of the challenges is the removal of radioactive fuel debris inside the reactor containment vessel [1]. In the experimental stage, grasping and suctioning are considered methods for removing debris [1]. Because suctioning is generally performed by a suction pump through a hose, the work of applying the hose to the target can be generalized as reaching. To protect operators from radioactive contamination, teleoperation, in which the human operator manipulates the construction equipment from a safe remote location, was used. However, the efficiency of teleoperation is approximately 50% lower than that of onsite work during excavations [2].

To improve efficiency, several studies have been conducted on visualizing the surrounding environment for operators in remote sites. Komatsu et al. [3] proposed an arbitrary viewpoint image generation system for the teleoperation of indoor robots using multiple fish-eye cameras and two-dimensional distance sensors. Sugasawa et al. [4] proposed a bird's-eye view generation method for dump trucks and hydraulic excavators using a fish-eye camera and a 3D distance sensor. Because the discussions focused on arbitrary viewpoints, progress in achieving the optimal viewpoint for specific tasks remains minimal. The choice of a viewpoint is essential when grasping by a robot arm. Gualtieri et al. [5] showed that choosing the correct viewpoint can significantly improve the average accuracy of the grasp.

In the field of unmanned construction, environmental cameras are often fixed. To reduce the blind spots caused by fixed viewpoints, Kamezaki et al. [6] proposed an autonomous camera control system using six displays. However, installing multiple environmental cameras at a disaster site is challenging because of the cost and effort involved. In addition, multi-display systems increase the cognitive load on the operator and require skill and experience for the operator to determine the optimal viewpoint [7]. Therefore, it is crucial to select a single optimal viewpoint. In a previous study on visualizing the surrounding environment in construction work, Chikushi et al. [8] proposed a method for automatically controlling external cameras based on the required specifications of construction machine operators in the construction of a dam. The methods in the aforementioned studies were evaluated by conducting subject experiments.

There has been much research on single optimal viewpoints for different tasks. Sato et al. [9] proposed the allowable range of single viewpoints and single optimal viewpoints for the digging and releasing tasks of a hydraulic excavator by conducting subject experiments.

However, subject experiments are limited because they cannot be conducted in advance when lacking information on special situations, such as inside the pedestal of a nuclear plant crushed in a disaster. On the contrary, utilizing computer simulation enables the investigation of the optimal viewpoint as soon as the information is acquired. However, because humans act based on past experiences, it becomes difficult to obtain an optimal viewpoint that is not affected by individual differences because the bias of each person's habits and preferences is included when the optimal viewpoint is investigated through questionnaires.

In this study, instead of using subject experiments, we extracted the factors that affect the optimal viewpoint using reinforcement learning, which is expected to replace subject experiments, thus dealing with the challenge of lacking environmental information in advance.

*A part of this study is financially supported by the Nuclear Energy Science & Technology and Human Resource Development Project (through concentrating wisdom) from the Japan Atomic Energy Agency / Collaborative Laboratories for Advanced Decommissioning Science.

[1]Department of Precision Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo, Tokyo 113-8656, Japan. {liuhaoxiang, komatsu, woo, yamashita, asama}@robot.t.u-tokyo.ac.jp

[2]Department of Robotics, Division of Mechanical Engineering, Tohoku University, 6-6-01, Aza-Aoba, Aramaki, Aoba-ku, Sendai-shi, Miyagi 980-8579, Japan. ytamura@tohoku.ac.jp

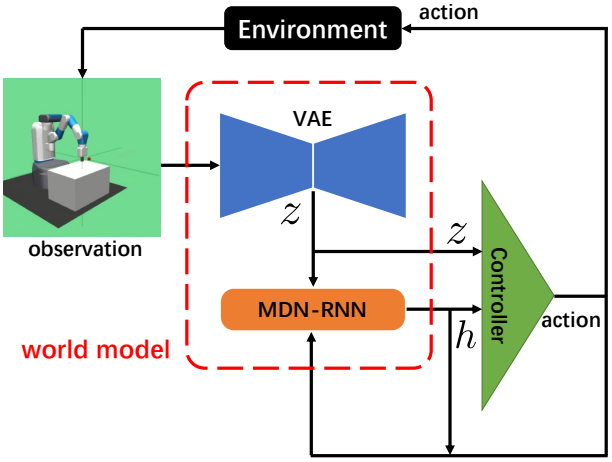


Fig. 1: Schematic of World Models.

II. Proposed Method

In this study, we employed the world model proposed by Ha et al. [10] using the PyTorch framework [11]. Using the reinforcement learning reward as an indicator, we extracted the elements that affect the optimal viewpoint. Compared to conventional methods, such as the subject experiment, the proposed method is not affected by the bias of each participant. Extracting an optimal viewpoint that is not affected by individual differences is expected.

To process large amounts of information, human brains learn information as abstract spatial and abstract temporal representations [12]. Humans are capable of observing and abstractly describing a scene. For example, when making decisions while driving, the brain does not analyze every pixel of the image information in the field of view. However, it transforms the visual information into low-dimensional representations, such as one's position on the road or curvature to determine the next action.

Figure 1 shows a schematic of the world model. In the world model, variational autoencoder (VAE) [13] [14] is used to transform the high-dimensional input observation into an abstract low-dimensional latent space. In this study, we take a three-channel 64 pixel 64 pixel image as input and convert it into a 32-dimensional latent vector. Then, to predict the future based on the current information, the MDN-RNN [15] [16] is used. Using mixture distribution networks (MDNs), the next latent vector is predicted from a mixture Gaussian distribution. Finally, the controller can return an action with the latent vector transformed by the VAE, representing the current state, and the latent vector output by the recurrent neural network (RNN) as input. Because there is no teaching data for the appropriate action in each scene, we need to train the controller using reinforcement learning.

1) VAE: By minimizing the reconstruction error and Kullback-Leibler (KL) divergence, VAE can transform a high-dimensional input image into a latent space approximating a multivariate standard normal distribution. As shown in Fig. 2, the VAE consists of an encoder and a decoder. The encoder uses a convolutional neural network to output the mean μ and the logarithm of the variance $\log\sigma^2$ of each of the 32-dimensional latent vectors. Subsequently, the latent vector z can be represented by sampling ϵ from the standard normal distribution as:

$$z = \mu + \epsilon \exp(\frac{\log \sigma^2}{2}), \tag{1}$$

where z is used as the input to the decoder, and the decoder returns the reconstructed image as output. Compared to the conventional Autoencoder (AE), the latent vectors in Eq. (1) are in the neighborhood of μ. Owing to the minimization term of the KL divergence, the decoder must ensure that neighboring points in the latent space reconstruct similar images to reduce the reconstruction error. The improvements above increase the chances that the decoder will reconstruct the correct image when a latent vector not previously in the training data is input to the decoder.

2) MDN-RNN: The structure of the MDN-RNN is shown in Fig. 3. It consists of a long short-term memory (LSTM) and a dense layer MDN that converts the latent vector of the LSTM into a mixture Gaussian distribution [17]. The input of the LSTM is 32 continuous 35-dimensional vectors, each composed of a 32-dimensional latent vector output by the VAE and a 3-dimensional action vector. The input is then transformed into a 256-dimensional output, and the dense layer transforms it into a 327-dimensional vector.

3) Controller: The structure of the network used in the controller is illustrated in Fig. 4. A 288-dimensional vector, which is a combination of the 32-dimensional latent vector of VAE and 256-dimensional latent vector of LSTM, is used as input, and the network outputs a three-dimensional action vector without any hidden layer. Because no teaching data indicates the robot arm's optimal motion, we use reinforcement learning to train the robot arm. The agent that learns the action of the robot arm causes trial and error in the environment based on the magnitude of the reward it receives. We define $\mathbf{x}_{\mathrm{robot}}$ as the tip of the robot arm and $\mathbf{x}_{\mathrm{target}}$ as the target's position. The reward function is represented by

$$r_i = -\|\mathbf{x}_{\mathrm{target}} - \mathbf{x}_{\mathrm{robot}}\|, \tag{2}$$

where the negative Euclidean distance between the tip of the robot arm and the target in each frame is given as a reward. Through all episodes (simulations), the tip of the robot arm, $\mathbf{x}_{\mathrm{robot}}$, starts from the same position. Because an episode (one simulation) is set to n frames, the sum of the rewards in each frame, r_{total}, is defined as follows:

$$r_{\mathrm{total}} = \sum_{i=1}^{n} r_i, \tag{3}$$

where n is set to 50 in this work.

We designed the reward function above for the following two reasons:

1) The closer the robot arm gets to the target object, the greater the reward.
2) The faster the robot arm reaches the target, the greater the reward.

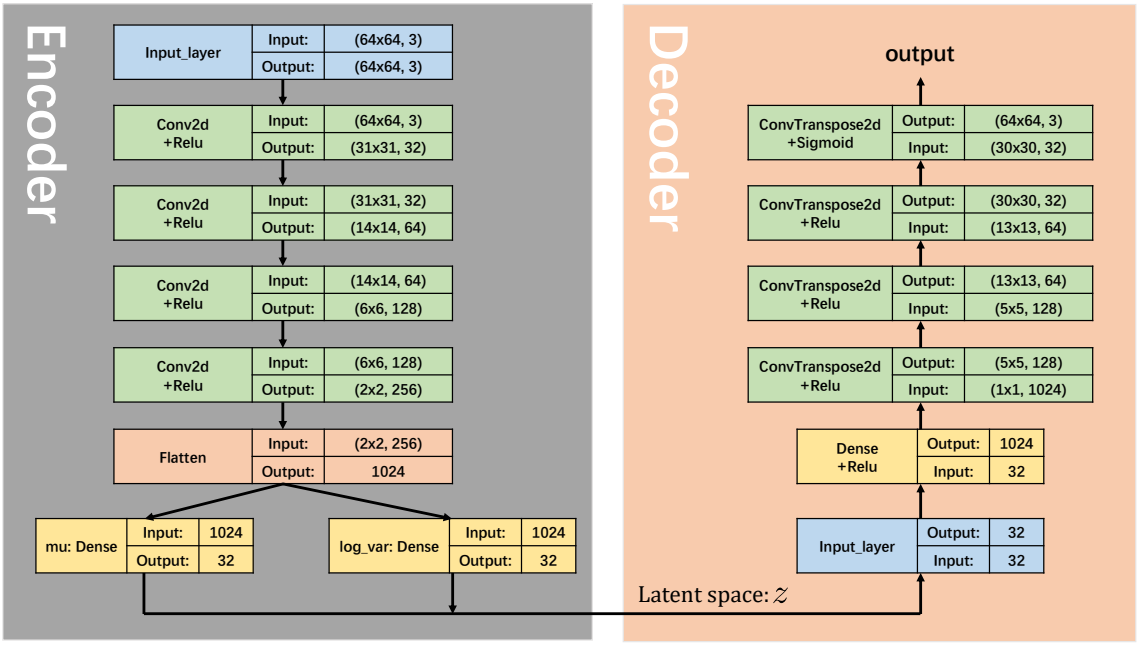


Fig. 2: Schematic of VAE. The VAE consists of an encoder and a decoder. The encoder uses a convolutional neural network to output the mean μ and the logarithm of the variance $\log\sigma^2$ of each of the 32-dimensional latent vectors. Subsequently, the latent vector z can be represented by Eq. (1). The decoder receives z as the input and returns the reconstructed image as output.

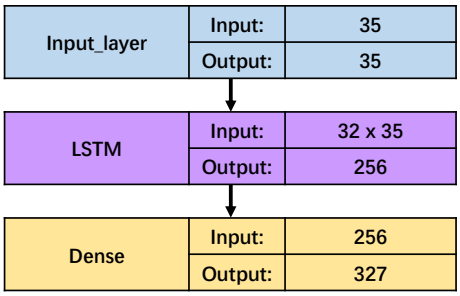


Fig. 3: Schematic of MDN-RNN. The input of the LSTM is 32 continuous 35-dimensional vectors, each composed of a 32-dimensional latent vector output by the VAE and a 3-dimensional action vector. The input is then transformed into a 256-dimensional output, and the dense layer transforms it into a 327-dimensional vector.

Therefore, we expect rapid and precise reaching when learning yields the biggest reward. Specifically, we use the covariance matrix adaptation evolution strategy (CMA-ES) [18] to train the controller. The CMA-ES process was as follows:

1) Create multiple agents and initialize the parameters to be optimized for each agent.
2) Loop through the following operations:
   (a) Evaluate each agent in the environment and return the average total reward for multiple episodes.
   (b) Breed the agent with the best score and generate a new agent, adding randomness to the parameters.
   (c) Add newly created agents and remove poorly performing agents.

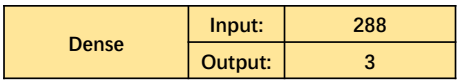


Fig. 4: Schematic of Controller. A combination of the 32-dimensional latent vector of the VAE and the 256-dimensional latent vector of LSTM is used as the input, and the output is a three-dimensional action vector.

III. Simulation and Results

We extracted six different viewpoints and ran the simulation four times under the same constraints. The polar coordinates of the body system presented in Fig. 5 were used to describe these viewpoints. The viewpoint angles used in the simulation are listed in Table I, and the number of epochs is listed in Table II. The training of VAE and RNN started with a learning rate of 0.001, and if the loss did not improve for five consecutive epochs, the learning rate was halved, and the training was continued. In addition, if the loss did not improve for more than 30 epochs, we considered that the model had stopped improving. Consequently, we stopped training to prevent overtraining and save time. We also set up four options for the location of the target and placed them at the corners of the table, as shown in Figs. 6(a), 6(c), 6(d), and 6(f), respectively. The simulation results for each viewpoint are presented in Figs. 7. The horizontal axis represents the epoch, and the vertical axis represents the total reward in

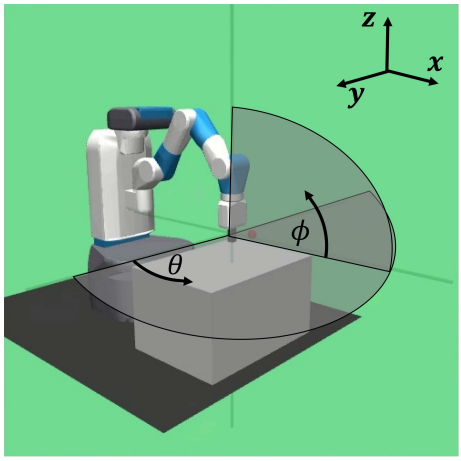


Fig. 5: Polar coordinates of the robot arm. θ is the pan angle, and ϕ is the tilt angle.

TABLE I: Selection of viewpoints

	a	b	c	d	e	f
θ [°]	0	45	90	0	45	90
ϕ [°]	30	30	30	45	45	45

the corresponding episode. The solid and dashed lines in the figure represent the mean values, and the semi-transparent region represents the 95% confidence interval.

A. Comparison of θ-direction

The controller was trained for 10,000 epochs, and the rewards evaluated every five epochs are shown in the graphs. The results for $\phi = 30°$ are shown in Fig. 8(a), where $\theta = 90°$ has the largest reward. The results for $\phi = 45°$ are shown in Fig. 8(b), where the differences between each θ angle are small.

B. Comparison of ϕ-direction

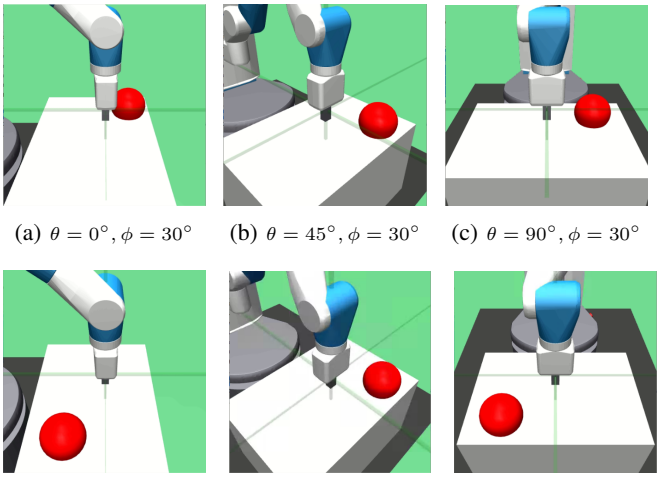
The controller was trained for 10,000 epochs, and the rewards evaluated every five epochs are shown in the graphs. The results for $\theta = 0°$, $\theta = 45°$, and $\theta = 90°$ are shown in Figs. 9(a), 9(b), and 9(c), respectively. Fig. 9 shows that the reward is maximized when $\phi = 30°$, except in the early stages of training.

C. TOP25% reward evaluation

Because rewards are recorded every five epochs for 10,000 epochs of training, we sorted the 2,000 data in order of increasing reward and retrieved the largest 500 rewards. Then, we calculated the average, the results of which are listed in Table III. The bold-type characters represent the higher reward in each column, where θ is fixed.

IV. Discussion

In this study, we used the reinforcement learning reward to demonstrate working efficiency. The greater the reward, the higher the efficiency. As shown in Table III, for the same θ, when ϕ is 30°, the reward is higher than when ϕ is 45°. The reason for this could be that the viewpoint moves on



(a) $\theta = 0°, \phi = 30°$ (b) $\theta = 45°, \phi = 30°$ (c) $\theta = 90°, \phi = 30°$

(d) $\theta = 0°, \phi = 45°$ (e) $\theta = 45°, \phi = 45°$ (f) $\theta = 90°, \phi = 45°$

Fig. 6: Viewpoints used for learning in Table I.

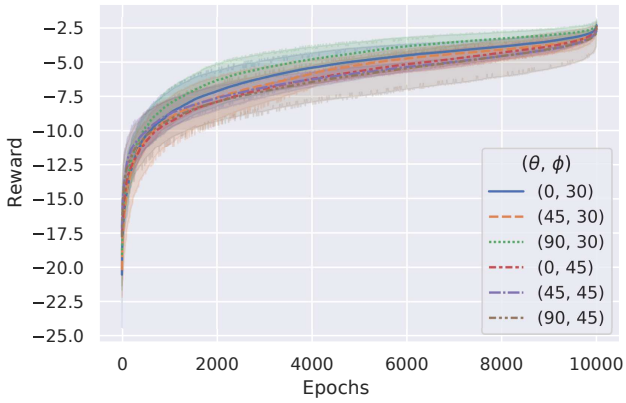


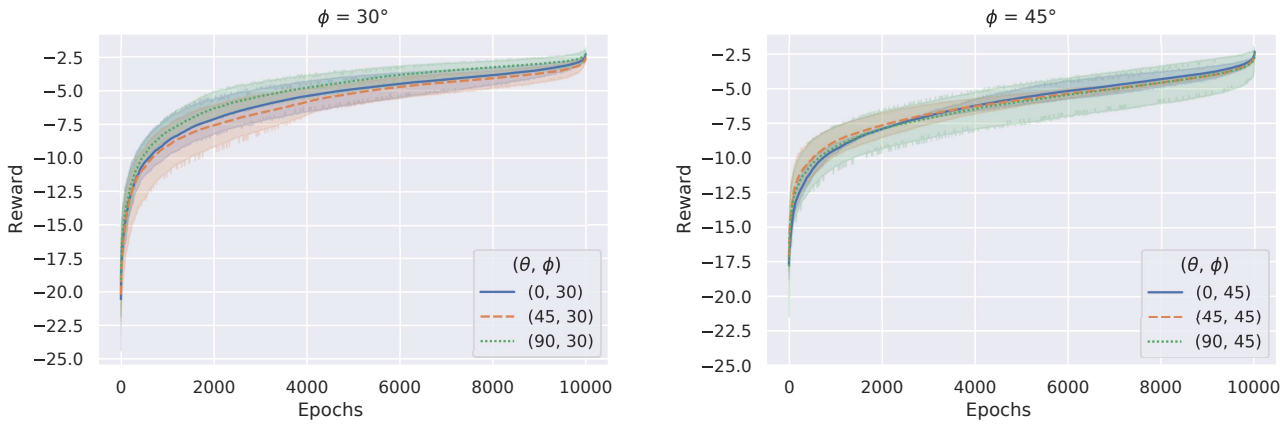
Fig. 7: All-viewpoints results. The horizontal axis represents the epoch, and the vertical axis represents the total reward in the corresponding episode. The solid and dashed lines in the figure represent the mean values, and the semi-transparent region represents the 95% confidence interval.

a sphere, as shown in Fig. 10, when the robot arm moves the same distance along the vertical direction, the change in pixels is more significant when ϕ is 30° than when ϕ is 45°. Specifically, the larger the angle ϕ of the viewpoint, the harder it is to perceive the vertical motion of the robot arm, which results in a lower reward.

The difference above caused by the tilt angle ϕ has also been observed in previous research [9]. Sato et al. used a subject experiment to investigate the optimal viewpoint of the skilled operator and mentioned that the lower the tilt angle ϕ in the release movement of the hydraulic excavator, the higher the efficiency of the operation because the operator knows the vertical distance to the ground.

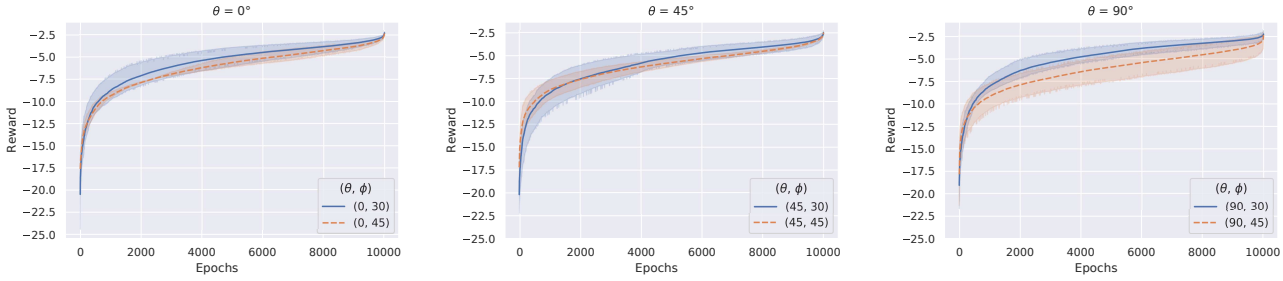
The statement above is the same as the result of this study, suggesting the validity of the world model in evaluating the proper viewpoint for human operators.

In addition, in a previous study [9], two different subjects



(a) $\phi = 30°$ (b) $\phi = 45°$

Fig. 8: Comparison of θ. We fixed ϕ on this occasion to contrast the differences caused by θ.



(a) $\theta = 0°$ (b) $\theta = 45°$ (c) $\theta = 90°$

Fig. 9: Comparison of ϕ. We fixed θ on this occasion to contrast the differences caused by ϕ.

TABLE II: The number of epochs

	Epochs
VAE	200
MDN-RNN	60
Controller	10,000

TABLE III: TOP25% reward evaluation

	$\theta = 0°$	$\theta = 45°$	$\theta = 90°$
$\phi = 30°$	**-3.52**	**-3.77**	**-3.06**
$\phi = 45°$	-3.89	-4.11	-3.83

took 17.0 seconds and 47.6 seconds, respectively, for the same task at the same viewpoint. The time differed significantly owing to individual differences. Conversely, the method proposed in this study can avoid the influence of individual differences using reinforcement learning.

V. Conclusion

In this study, we used a world model [10] and reinforcement learning to evaluate the optimal viewpoint for work efficiency. The simulations of multiple viewpoints showed that the world model demonstrated a trend similar to that of the human subject experiment. In computational neuroscience, the manifold hypothesis posits that the distribution

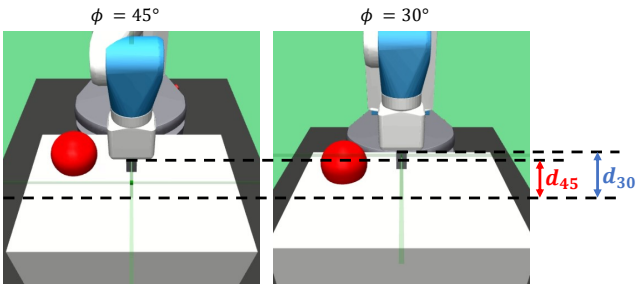


Fig. 10: Comparison of viewpoints in $\theta = 90°$. The viewpoint moves on a sphere; when the robot arm moves the same distance along the vertical direction, the change in pixels is larger when ϕ is 30° than when ϕ is 45°.

of data observed in the real world can be perceived as a low-dimensional manifold. Just as the human brain constantly converts high-dimensional information into abstract low-dimensional information, the VAE of the world model performs the same function. Our simulation results suggest the possibility of evaluating the optimal viewpoint for humans using the world model. The use of reinforcement learning instead of subject experiments has the advantage of dealing with the lack of environmental information in advance and eliminating individual differences in subject experiments.

ACKNOWLEDGMENT

This work was supported by JAEA Nuclear Energy ST and Human Resource Development Project Grant Number-JPJA19H19210047.

REFERENCES

[1] Tokyo Electric Power Company Holdings, Inc.: "Mid-and-Long-Term Decommissioning Action Plan 2020", March 27. 2020, https://www.tepco.co.jp/en/hd/decommission/information/dap/pdf/dap_20200327_01-e.pdf, accessed July 9. 2021.

[2] K. Chayama, A. Fujioka, K. Kawashima, H. Yamamoto, Y. Nitta, C. Ueki, A. Yamashita, and H. Asama: "Technology of Unmanned Construction System in Japan", *J. Robot. Mechatron.*, Vol. 26, No. 4, pp. 403–417, 2014.

[3] R. Komatsu, H. Fujii, Y. Tamura, A. Yamashita, and H. Asama: "Free Viewpoint Image Generation System Using Fisheye Cameras and a Laser Rangefinder for Indoor Robot Teleoperation", *ROBOMECH Journal*, Vol. 7, No. 1, pp. 1–10, 2020.

[4] Y. Sugasawa, S. Chikushi, R. Komatsu, J. Y. Louhi Kasahara, S. Pathak, R. Yajima, S. Hamasaki, K. Nagatani, T. Chiba, K. Chayama, A. Yamashita, and H. Asama: "Visualization of Dump Truck and Excavator in Bird's-eye View by Fisheye Cameras and 3D Range Sensor", *Proceedings of the 16th International Conference on Intelligent Autonomous Systems (IAS)*, pp. 480–491, 2021.

[5] M. Gualtieri and R. Platt: "Viewpoint Selection for Grasp Detection", *Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 258–264, 2017.

[6] M. Kamezaki, J. Yang, H. Iwata, and S. Sugano: "An Autonomous Multi-camera Control System Using Situation-based Role Assignment for Tele-Operated Work Machines", *Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5971–5976, 2014.

[7] D. E. Crundall and G. Underwood: "Effects of Experience and Processing Demands on Visual Information Acquisition in Drivers", *Ergonomics*, Vol. 41, No. 4, pp. 448–458, 1998.

[8] S. Chikushi, Y. Moriyama, H. Fujii, Y. Tamura, H. Yamakawa, K. Nagatani, Y. Sakai, T. Chiba, S. Yamamoto, K. Chayama, A. Yamashita and H. Asama: "Automated Image Presentation for Backhoe Embankment Construction in Unmanned Construction Site", *Proceedings of the 2020 IEEE/SICE International Symposium on System Integration (SII)*, pp. 22–27, 2020.

[9] R. Sato, M. Kamezaki, M. Yamada, T. Hashimoto, S. Sugano, and H. Iwata: "Environmental Camera Placements for Skilled Operators in Unmanned Construction", *Automation in Construction*, Vol. 119, p. 103294, 2020.

[10] D. Ha and J. Schmidhuber: "World Models", *arXiv preprint arXiv:1803.10122*, 2018.

[11] A. Paszke, et al.: "Pytorch: an imperative style, high-performance deep learning library", *Advances in Neural Information Processing Systems*, Vol. 32, pp. 8024–8035, 2019.

[12] L. Chang and D. Y. Tsao: "The Code for Facial Identity in the Primate Brain", *Cell*, Vol. 169, No. 6, pp. 1013-1028, 2017.

[13] D. P. Kingma and M. Welling: "Auto-encoding Variational Bayes", *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[14] D. J. Rezende, S. Mohamed, and D. Wierstra: "Stochastic Backpropagation and Approximate Inference in Deep Generative Models", *Proceedings of the International Conference on Machine Learning*, pp. 1278–1286, 2014.

[15] C. M. Bishop: "Mixture Density Networks", *Technical Report*, Aston University, 1994.

[16] A. Graves: "Generating Sequences with Recurrent Neural Networks", *arXiv preprint arXiv:1308.0850*, 2013.

[17] S. Hochreiter and J. Schmidhuber: "Long Short-Term Memory", *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

[18] N. Hansen: "The CMA Evolution Strategy: A tutorial", *arXiv preprint arXiv:1604.00772*, 2016.