2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Real-time Coordinated Motion Generation: A Hierarchical Deep Predictive Learning Model for Bimanual Tasks

Genki Shikada[1] and Simon Armleder[2] and Hiroshi Ito[1] and Gordon Cheng[2] and Tetsuya Ogata[1,3]

Abstract— **Robots that autonomously operate in human living environments require the ability to adapt to unpredictable changes and flexibly handle a variety of tasks. Particularly, coordinated bimanual motions are essential for enabling tasks that are difficult with just one hand, such as grasping bulky objects, transporting heavy loads, and precision work. Traditional methods of generating robot motions typically involve executing pre-programmed motions, making it challenging to adapt to complex and unpredictable environmental changes. To address this issue, our research focuses on generating diverse motions that can flexibly adapt to environmental changes based on Deep Predictive Learning from a small amount of real-world data. Previous Deep Predictive Learning models have generated the motions of a robot's left and right arms by a single LSTM, making it difficult to operate them independently. Therefore, we propose a new Hierarchical Deep Predictive Learning model specialized for generating coordinated bimanual motions. This model comprises three components: a Left-LSTM, which learns the body and visual information on the robot's left side, a Right-LSTM that performs a similar function for the right side, and a Union-LSTM which integrates this information at a higher level. To verify the effectiveness of the proposed model, we conducted bimanual grasping experiments with multiple different objects using two different robots. The experimental results showed that independent of hardware, our model demonstrated a higher success rate compared to the traditional approach, indicating its enhanced capability in coordinating bimanual motions.**

I. INTRODUCTION

In recent years, there has been an increasing demand for robots capable of autonomously operating in human living environments and assisting people in their daily lives. Reaching this goal will require robots that can adapt to unpredictable changes and handle a wide range of tasks. Coordinated bimanual motions, in particular, are important in scenarios closely related to humans, such as support within households and tasks in factories, as they enable actions that are difficult with just one hand, like grasping bulky objects, transporting heavy loads and precision work [1].

Traditional methods of generating robot motions typically involve replaying motions pre-set through programming, which is effective for predictable environments and repetitive tasks, but struggles with complex and unpredictable situations. Deep learning-based robot motion generation has been gaining attention as a solution to this challenge. By utilizing deep learning, it becomes possible to directly process complex sensor data from the environment and generate adaptive motions capable of handling unknown situations.

However, collecting vast amounts of training data for deep reinforcement learning approaches is still very expensive. For instance, [2] required 14 robots to operate continuously for 2 months to learn object grasping. Therefore, training usually happens in simulation where data collection is cheap. However, the sim-to-real gap makes transferring these learned policies into the real world challenging [3], [4].

One solution to this problem is a method called Deep Predictive Learning [7], which enables the generation of diverse motions that adapt to environmental changes from a small amount of data collected in the real world. Deep Predictive Learning models learn the temporal relationship between the robot's sensory data and motion data obtained during human demonstrations. During task execution, they predict both, the expected sensation and the next motor command in real-time from the current sensorimotor information. The predicted motor command is then sent to the robot's low-level controller, achieving adaptive behaviors where the robot reacts to changes in its environment.

In this study, we adopt a hierarchical structure that is specifically designed to generate coordinated bimanual motion. Our method uses two low-level LSTMs for the left and right arms of the robot. Each has access to the visual and proprioceptive signals of its respective side. The internal states of this lower layer are then further fused inside a high-level LSTM that allows for coordinated motions between the two arms. By first, separating the sensorimotor information on the left and right sides, the model can generate independent motions for the two arms. This is, for example, required when reaching for different points on an object before picking it up. At a higher level, the Union-LSTM couples the motion of both arms to synchronize them when needed. For example, to lift an object at the same time.

To verify the effectiveness of the proposed model, we conducted bimanual grasping experiments with multiple different objects using two robots. The experimental results showed that our model demonstrated a higher success rate compared to traditional models that predict the motions of both arms together, on both robots. This indicates that the proposed model is better suited for bimanual robots that require coordinated motion between both arms.

[1]Genki Shikada, Hiroshi Ito and Tetsuya Ogata are with the Department of Intermedia Art and Science, Waseda University, Tokyo, Japan shikashika@fuji.waseda.jp

[2]Simon Armleder and Gordon Cheng are Institute for Cognitive System (ICS) Technical University of Munich 80333, Germany

[3] Tetsuya Ogata is with the National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan ogata@waseda.jp

II. RELATED WORKS

A. Deep Predictive Learning

Deep Predictive Learning is a framework inspired by the mechanism through which the human brain infers the causes of sensory inputs and learns, through the minimization of free energy, a concept from statistical physics [5]. In this approach, Recurrent Neural Networks (RNN) are used to learn how to minimize the prediction error of sensorimotor information between the current and next steps. During learning, the robot learns the sensorimotor information from its previous actual motions and acquires a data-driven motion generation model. This learning data does not require labels, and the model is trained to predict the robot's state in the next step using its current state as input. This autoregressive learning eliminates the need for designing environmental physical models traditionally required in robotics, significantly reducing the costs associated with robot motion generation. During motion execution, the robot predicts the near-future sensory and motor states in real time based on sensorimotor information and performs motions to minimize the error between reality and prediction. This ability to flexibly adjust the differences between learning and execution in real-time enables adaptive motions in unknown environments.

As examples of robot motion generation using Deep Predictive Learning, studies such as fabric folding by Yang et al. [6], door opening by Ito et al. [7], chest opening by Ichiwara et al. [8], knot tying by Suzuki et al. [9], and cooking with tools by Saito et al. [10] have been cited. These studies demonstrate cost-effective, real-world motion generation adapted to changes in the surrounding environment by combining Convolutional Neural Network (CNN)-based image recognition with RNN-based robot motion prediction. In the tasks by Yang et al. [6], Ichiwara et al. [8], and Suzuki et al. [9], robust motion generation has been demonstrated for flexible objects, whose shapes constantly change and for which environmental modeling is challenging. Particularly, Ichiwara et al. [8] improved the robustness to changes in object states by using a spatial attention mechanism in CNN-based image recognition, which focuses on task-relevant positional information from the image.

B. Dual arm robot motion generation

The realization of coordinated bimanual motions plays a significant role in the advancement of robotics technology, and there is active research on dual-arm robots and methods for generating bimanual motions. Regarding the hardware, Diftler et al. [11] developed a bimanual robot capable of performing tasks using the same hardware and interface as humans, enabling the robot to carry out experimental tasks on the space station. Regarding deep learning motion generation, Zipeng et al. [12] developed "Mobile ALOHA", a low-cost solution for collecting learning data for mobility and bimanual tasks in practical daily environments, aimed at tasks requiring coordinated whole-body and dexterous manipulation. Additionally, as examples of bimanual motion generation using Deep Predictive Learning, fabric folding by Yang et al. [6] and knot tying by Suzuki et al. [9] are notable. These studies highlight the importance of bimanual motions in daily life and complex tasks, yet a detailed examination of the methods for achieving coordinated bimanual motions remains to be further explored.

To realize the coordinated bimanual motions for robots, human knowledge that "information from the right side of the body is transmitted to the left brain, and information from the left side of the body is transmitted to the right brain, where these pieces of information are integrated through the corpus callosum [13]" can be a reference. This process highlights the importance of the left and right arms referencing each other's information and working together in coordination. Applying this insight, Rakital et al. [15] conducted research that enables coordinated bimanual motions by utilizing linguistic information to execute complex tasks. Furthermore, applying this knowledge to the structure of deep learning models, Liu et al. [14] achieved coordinated bimanual movements in multi-agent deep reinforcement learning by treating the two robotic arms as independent agents and combining a reward system that encourages cooperative tasks with penalties that suppress competition. In this study, we apply these insights to the framework of Deep Predictive Learning and discuss the Hierarchical Deep Predictive Learning Model.

III. PROPOSED METHOD

A. Design concept

Human brains consist of long-term and short-term memory, and according to computational neuroscience research [16], [17], long-term memory is believed to hold broad plans of motion (for example, coordination between the left and right arms or the order of tasks), while short-term memory retains specific motor elements (such as stretching an arm or grasping an object).

This memory structure of the brain is also reflected in deep learning approaches. The concepts of long-term and short-term memory play a significant role in LSTM [18] and MTRNN (Multiple Timescales Recurrent Neural Network) [19]. The cell state in LSTM and the slow context layer in MTRNN correspond to long-term memory, while the hidden state in LSTM and the fast context layer in MTRNN serve the function of short-term memory. These models have been applied in the context of robot motion generation as well [8], [10]. In this case, primitive motion information is embedded in the hidden states or fast context layers, while the cell states or slow context layers embed the "know-how" of motions, including long-term dependencies of motions, context of motions patterns, and motion strategies. Based on the insights from the human brain and deep learning models mentioned above, this paper proposes a model that integrates the cell states of LSTM to realize coordinated bimanual motions in robots.

B. Model architecture

In this paper, we present a Deep Predictive Learning model designed to achieve coordinated bimanual motions in robots (Fig.1). The model is a hierarchical neural network

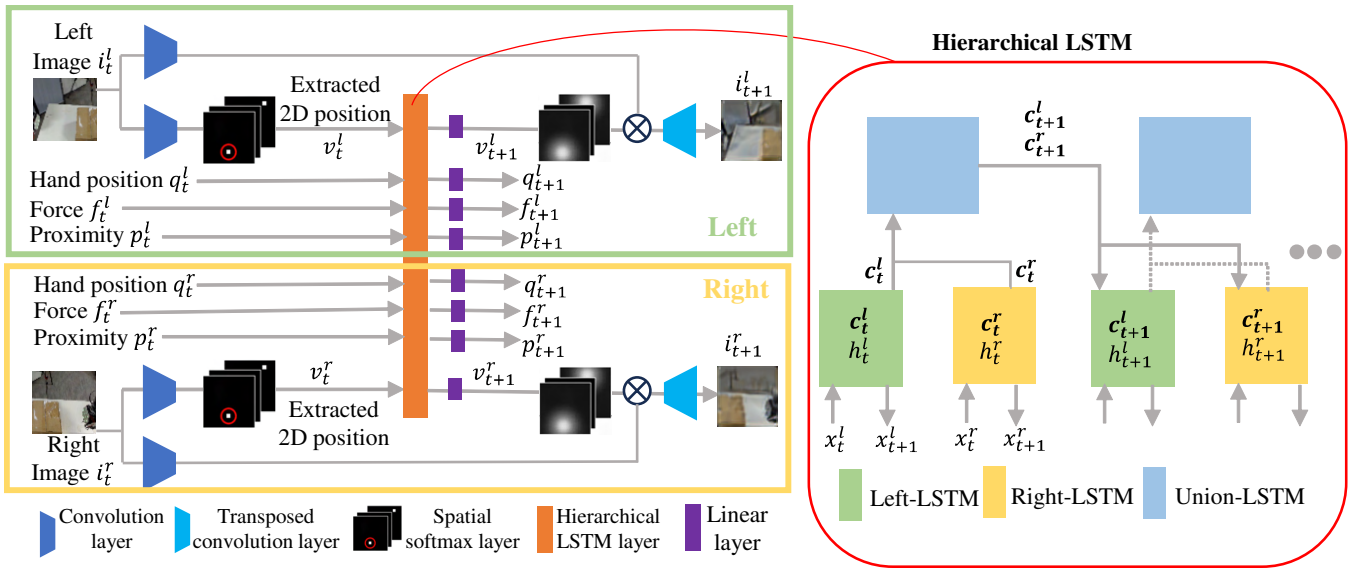

Fig. 1. The proposed Hierarchical LSTM model for bimanual coordinated motion. It consists of three parts: Left-LSTM for predicting the motion of the left arm, Right-LSTM for predicting the motion of the right arm, and Union-LSTM for integrating these. Left-LSTM and Right-LSTM take as inputs attention points extracted from images, haptic information, and motion information.

composed of a Left-LSTM for learning the body and visual information of the robot's left side, a Right-LSTM for similarly learning the right side's information, and a Union-LSTM for integrating this information at a higher level. Based on the SARNN (Spatial Attention Recurrent Neural Network) approach [8], referenced from the literature, both the Left-LSTM and Right-LSTM learn the temporal relationships between each side's body and visual information. SARNN utilizes a CNN layer and a Spatial softmax layer to explicitly extract important positional information from input images and learns the temporal relationship with the robot's body information, generating robust motions in response to changes in the position of objects. Coordinated bimanual motions are generated by combining the internal states of the Left-LSTM and Right-LSTM and integrating them through the Union-LSTM.

The robot's motor sensory information includes the motion information of the left and right arms such as joint angles and hand coordinates (q_t), the visual attention points (v_t) extracted from a camera, and the tactile information about hand contacts such as force (f_t) and proximity (p_t). At each step, the left and right LSTM predict the next sensory-motor signals based on the current sensory information (Eq. 1). The cell states (c_t) of these two low-level LSTMs are updated by the Union-LSTM (Eq. 2).

$$\begin{cases} s_t = [v_t, q_t, f_t, p_t] \\ \hat{s}_{t+1}^l = LSTM_{Left}(s_t^l, h_{t-1}^l), \\ \hat{s}_{t+1}^r = LSTM_{Left}(s_t^l, h_{t-1}^r), \end{cases} \tag{1}$$

$$\begin{cases} x_t^{union} = Concatenate(c_t^{left}, c_t^{right}) \\ \hat{x}_{t+1}^{union}, h_t^{union} = LSTM_{Union}(x_t^{union}, h_{t-1}^{union}), \\ \hat{c}_{t+1}^{left}, \hat{c}_{t+1}^{right} = Split(\hat{x}_{t+1}^{union}) \end{cases} \tag{2}$$

C. Learning

For preprocessing, we normalized the robot's body information to the range [0.0, 1.0]. We also set the number of intermediate nodes to 50 for both the Left-LSTM and Right-LSTM, and to 20 for the Union-LSTM. We employed the Adam optimization algorithm with the Leaky ReLU activation function, and we trained the model for 1000 epochs.

D. Real-time motion generation

Real-time motion generation is performed by inputting the visual information, tactile information, and motion information (joint angles or hand position) at the current time (t), and predicting their values for the next time ($t+1$). The inference cycle was set to 10 Hz, and the initial value of the hidden state (h_t) was set to 0.

IV. EXPERIMENT 1: TOMM

A. Experiment setting

We conducted bimanual object-grasping experiments to demonstrate that the proposed hierarchical LSTM model (Fig. 1) can perform coordinated bimanual tasks more effectively than conventional single LSTM models (Fig. 4) [8], [10]. We taught grasping motions for objects of various sizes and used the training data with a certain level of diversity to train the model, then verified whether the model could appropriately generate motions for objects of unknown sizes. Moreover, experiments were conducted using two robots to show that the proposed model is not dependent on specific hardware.

In the first experiment, the robot uses both arms to lift large cardboard boxes of various sizes. The box dimensions are too large to be grasped with a single robot hand. And then, we used the Tactile Omnidirectional Robot Manipulator

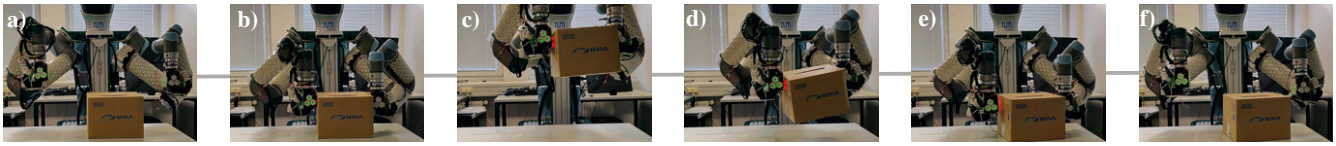

Fig. 2. Snapshot of TOMM's real-time motion generation. a) & b) is Reaching motion, c) & d) is Lift-up motion, e) & f) is Putting motion

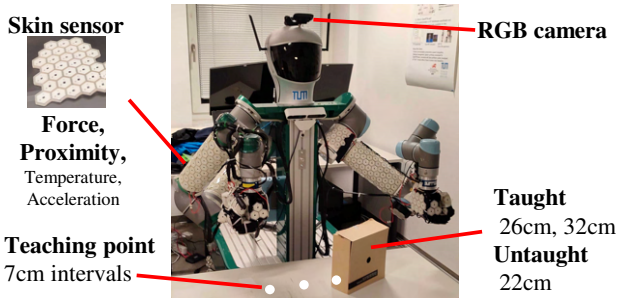

Fig. 3. TOMM and experiment environment. The box dimensions are too large to be grasped by the robot's hand. As such the robot requires both hands to lift the bulky box.

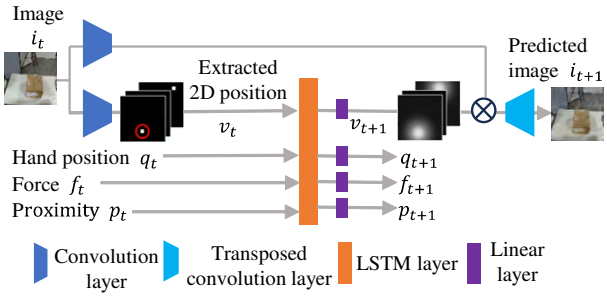

Fig. 4. Traditional single LSTM model. This model predicts the sensory and motion information of both the left and right arms in one LSTM.

(TOMM) [21] owned by the Institute for Cognitive Systems (ICS) at the Technical University of Munich. TOMM's arms consist of two UR5 arms, with Allegro Hands mounted at the end effectors. Both the arms and the hands are covered with skin sensors [22] capable of measuring force, proximity, acceleration, and temperature. In this study, we acquired force and proximity information from the end effectors and used them as contact information when the robot touches an object. A RealSense camera was attached to the top of the robot's head and provides visual feedback. The setup of TOMM and experimental environment is shown in Fig. 3.

To generate learning data for the model in Experiment 1, we taught the robot to grasp two different sizes of cardboard boxes (26 cm and 32 cm in width) using both hands through a teleoperation system based on HTC Vive. The cardboard boxes were placed in three different positions, and four teaching sessions were conducted at each position, collecting a total of 24 sets of teaching data. For each data set, we acquired images (64 $\times$ 64 $\times$ 3[pixel] $\times$ 2) for each arm, hand position coordinates (3-DoF $\times$ 2), force sensor values (1-DoF $\times$ 2), and proximity sensor values (1-DoF $\times$ 2) at a sampling rate of 10Hz over 18 seconds, using a total of 180 steps of time-series data. The camera images were trimmed so that the image for the left-LSTM included the left end of the object and the left arm, and similarly for the right-LSTM.

The extracted attention point coordinates from the camera images (v_t), hand positions (q_t), force sensor values (f_t), and proximity sensor values (p_t) were used as inputs to the LSTM. Based on this information, the model predicts the future sensory-motor signals (Eq.1, 2) of which the hand positions are sent as the next command to the robot controller.

B. Experiment result

Fig. 2 shows the real-time motion generation process using the proposed model. The motion sequence consists of recognizing the cardboard box and extending the arms (Fig. 2 a), b)), grasping and lifting the box (Fig. 2 c), d)), and finally placing the box back on the table (Fig. 2 e), f)). The model successfully recognizes both sides of the cardboard and generates appropriate independent reaching motion for each arm. During lifting, the motion becomes synchronized and both arms move upward at the same speed, all while maintaining an appropriate grasping force.

Table I shows the number of successful grasping motions on cardboard boxes of untaught widths for both the proposed hierarchical model for dual arms and the traditional model that predicts the motions of both arms together. Both models were trained on the same data for the same amount of time and were tested on a cardboard box with untaught dimensions (22 cm in width). In Table I, "success" is defined as the situation where the robot grasps the object with both hands, lifts it, and then places it back at its original location. "Force error" occurs when the robot's low-level controller stops due to excessive force applied to the hand during the motion, and "fail" is when the grasping with both hands fails. The proposed hierarchical model for dual arms successfully completed the reaching motion 8 out of 10 times and completed the object grasping task 5 times. On the other hand, the traditional model that predicts both arms together successfully completed the reaching motion only 2 out of 10 times and failed to complete the task any time.

C. Attention point in real-time motion generation

Attention points play a crucial role in explicitly extracting important positional information from input images and generating robust motions in response to changes in the target

TABLE I
COMPARISON OF PROPOSED HIERARCHICAL LSTM AND TRADITIONAL SINGLE LSTM MODELS

model	success	force error	fail
Proposed (Hierarchical LSTM)	5/10	3/10	2/10
Traditional (Single LSTM)	0/10	2/10	8/10

object. In the case of grasping a cardboard box using two arms, the critical positional information includes both hands and the ends of the cardboard box, which serve as the target when reaching for the object prior to lifting it. It is desirable for attention points to appear in these important locations.

Fig. 5 shows the behavior of attention points during real-time motion. In the proposed model, attention points appear on both sides of the object of interest and both hands, indicating that the model has learned the crucial elements within the images for the task. In contrast, in the traditional model, attention points only appear on the right hand and the box, suggesting insufficient learning of the temporal relationship between the images and motion information of the left hand.

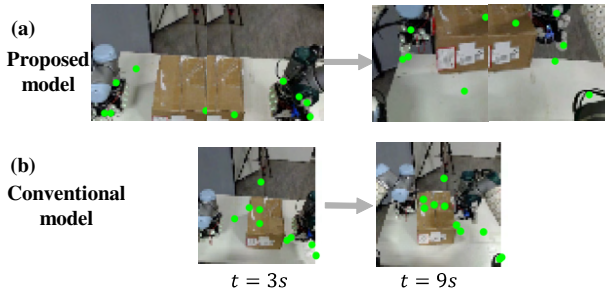

Fig. 5. Attention Point in real-time motion generation. (a) In the proposed hierarchical LSTM model, attention points appear on both arms and the target object. (b) In a traditional single LSTM model, attention points only appear on the right arm and the target object.

D. Force prediction

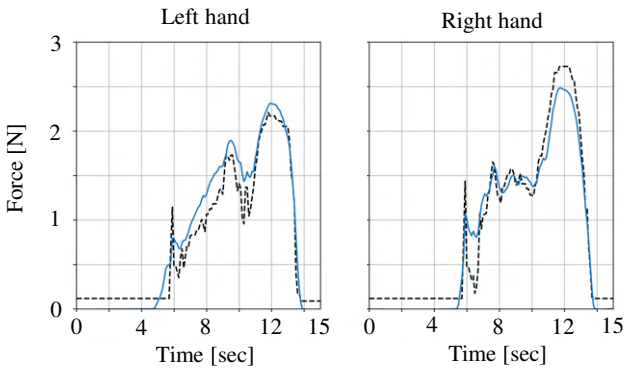

Fig. 6. Force sensor plot in real-time motion generation. The solid blue line represents the predicted values from the LSTM, while the dotted line represents the actual sensor values.

Bimanual cardboard grasping is classified as a tightly-symmetric motion according to Krebs et al.'s bimanual task classification [20]. In bimanual cardboard grasping motion, the arms operate symmetrically around the cardboard, requiring precise motion and force predictions to avoid dropping the box. Figure 6 shows the force sensor predictions made by the proposed LSTM model during real-time operation (solid blue line) and the actual force sensor values (dotted line). While the model tends to predict smoothly in response to the noise and sudden changes in the measured values, leading to some differences between the predicted and actual values, the model accurately captures the timing of increases and decreases in the measured values. This indicates that the proposed model is making appropriate force predictions.

E. LSTM state analyze

In robot motion generation using Deep Predictive Learning, robot motions are embedded in the LSTM's internal states. This latent representation is self-organized during the entrainment process [23]. For the cardboard grasping experiment, a model that has been appropriately trained is expected to form different attractors for motions corresponding to different cardboard boxes, while motions for the same cardboard box are expected to form similar attractors.

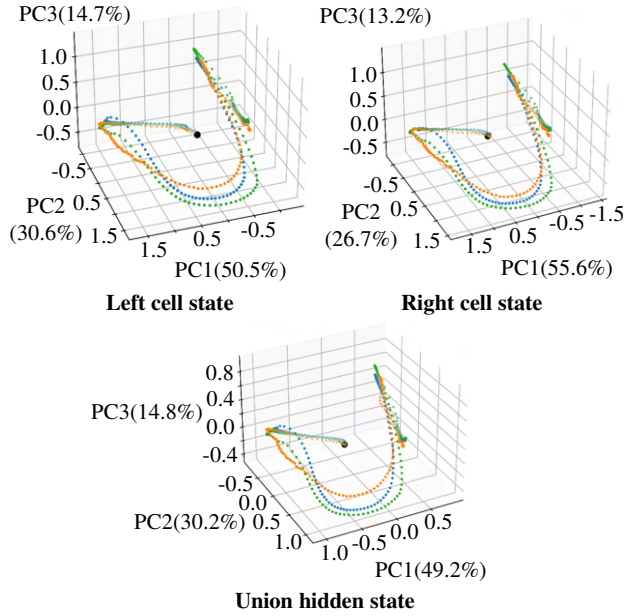

Fig. 7. Internal states of the proposed Hierarchical LSTM model. The orange is taught cardboard (32 cm), the blue is taught cardboard (22 cm), and the green is untaught cardboard (22 cm). Trajectories of the attractors are arranged in order of cardboard size.

Fig. 7 presents the visualization results of the internal states of the Left, Right, and Union-LSTM in the proposed model during real-time motion generation, using principal component analysis. The orange points represent the taught cardboard of 32 cm, the blue points represent the taught cardboard of 26 cm, and the green points represent the untaught cardboard of 22 cm. In the internal states of all LSTMs, attractors are formed in the order of the cardboard sizes, indicating that the proposed model acquired motions for

cardboard sizes extrapolated through learning. Furthermore, it was demonstrated that the arrangement of self-organized attractors in the lower layers of the Left-LSTM and Right-LSTM is appropriately learned and self-organized in the upper layer of the Union-LSTM as well.

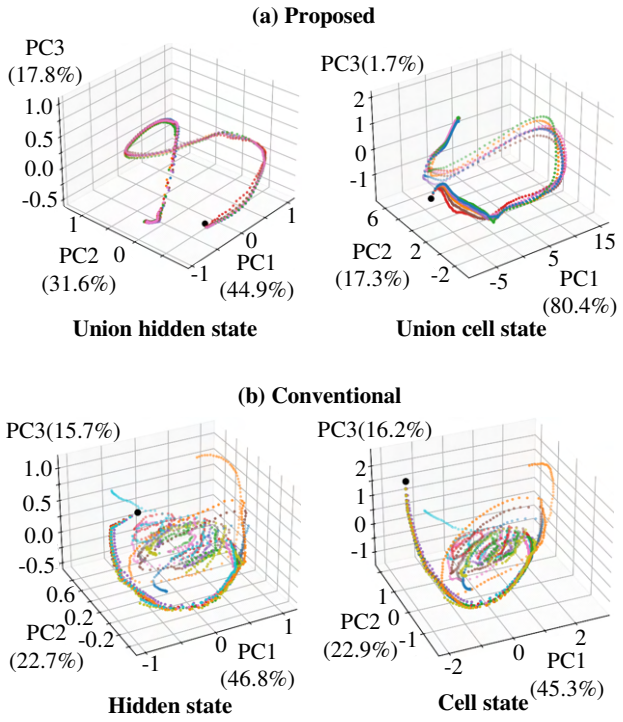

Fig. 8. Comparison of LSTM state of proposed hierarchical LSTM and conventional single LSTM model. (a) Proposed Method: Trajectories are converging. (b) Conventional Method: Trajectories are scattered.

Fig. 8 presents the visualization results, using principal component analysis, of the internal states during real-time motion generation for grasping untaught objects as shown in Table I, lists the success rates of the proposed model and the single LSTM model. In the successful cases of the proposed model, the internal states converge to a specific attractor, indicating that successful patterns are self-organized internally. On the other hand, in the single LSTM model, the attractors are scattered, indicating a lack of appropriate self-organization. The analysis of the internal states also shows that the traditional model does not generalize well to untaught cardboard dimensions and can't generate motions for new unseen objects.

V. EXPERIMENT 2: AIREC

A. Experiment setting

In the second experiment, we employ the proposed hierarchical model on a different bimanual robot. The task is to grasp four different-sized objects. Again, each of them is too large to be grasped with a single hand.

The robot used in Experiment 2 is the AIREC (AI-driven Robot for Embrace and Care), a dual-arm humanoid robot manufactured by Tokyo Robotics. AIREC features arms with

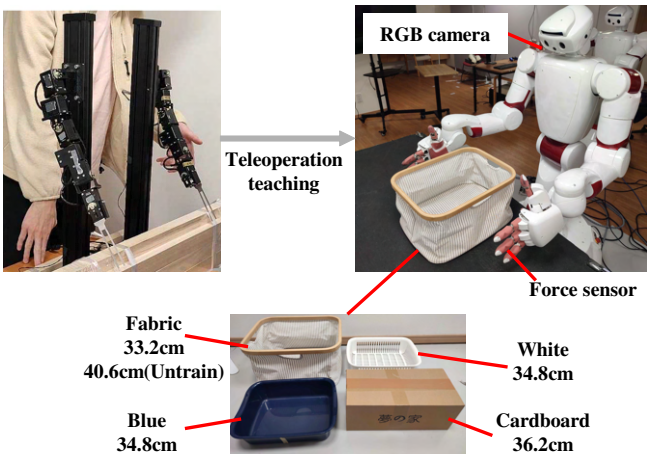

Fig. 9. AIREC and experiment environment

7-Dof hands with 6-Dof (3 for the thumb and 1 for each of the three fingers), equipped with pressure sensors, allowing for various object manipulations using force information. In this study, fingertip force information was acquired and used as contact information when the robot touched an object.

To generate learning data for the model, motions for grasping four different types of objects with both hands were taught using a remote control with a lead-through device. The four taught objects are a fabric box (33.2 cm), a cardboard box (36.2 cm), a blue plastic box (34.8 cm), and a white plastic box (34.8 cm) (Fig. 9). Each object was placed in three different positions, and four teaching sessions were conducted at each position, collecting a total of 48 sets of teaching data.

For each data set, images (64 $\times$ 64 $\times$ 3[pixel] $\times$ 2) for each arm, arm joint angles (6-DoF $\times$ 2), fingertip force sensor values for the index finger (4-DoF $\times$ 2), and fingertip force sensor values for the middle finger (4-DoF $\times$ 2) were acquired at a sampling rate of 10Hz over 15.1 seconds, using a total of 151 steps of time-series data. The camera images were trimmed so that the image for the left arm included the left end of the object and the left arm, and similarly for the right arm. The setup of AIREC and the experimental environment is shown in Fig. 9.

In Experiment 2, the extracted attention point coordinates (v_t), AIREC's arm joint angles (q_t), and fingertip force sensor values for the index and middle fingers (f_t) were used as inputs to the LSTM.

During real-time motion generation, the predicted joint angles for the next time step are inferred at 10 Hz by the Deep Predictive Learning model. For the actual motion command values, these predictions were upsampled to 50 Hz using linear interpolation (Eq. 3).

$$q(t) = (q1 - q0) \cdot \min\left(1.0, \max\left(0.0, \frac{t}{0.1}\right)\right) + q0 \quad (3)$$

where $q(t)$ is the interpolated value at the current time t, $q0$ is the starting value, and $q1$ is the ending value.

Fig. 10. Snapshot of AIREC's real-time motion generation. a) & b) is Reaching motion, c) & d) is Lift-up motion, e) & f) is Putting motion

B. Experiment result

TABLE II
COMPARISON OF PROPOSED HIERARCHICAL LSTM AND TRADITIONAL SINGLE LSTM MODELS

model	object	success	fall	fail
Proposed (Hierarchical LSTM)	Untrain	8/10	0/10	2/10
	Cardboard	4/10	4/10	2/10
	Blue	7/10	0/10	3/10
	White	10/10	0/10	0/10
Traditional (Single LSTM)	Untrain	7/10	0/10	3/10
	Cardboard	0/10	0/10	10/10
	Blue	0/10	0/10	10/10
	White	0/10	0/10	10/10

Fig. 2 illustrates the process of real-time motion generation using the proposed model. Table II shows the number of successful grasping motions for the proposed hierarchical dual-arm model and the traditional model that predicts the motions of both arms together. In Table II, "success" refers to successfully grasping the object with both hands, lifting it, and then placing it back in the original location. "Fall" means the object was grasped with both hands but was dropped during the process, and "fail" indicates a failure to grasp with both hands. "Untrain" represents the results of grasping a fabric box (40.6 cm) rotated 90 degrees from its orientation during teaching, "Cardboard" refers to a cardboard box (36.2 cm) in the same orientation as during teaching, "Blue" is a blue plastic box (34.8 cm), and "White" is a white plastic box (34.8 cm).

The traditional model failed 100% of the time in grasping the three types of boxes, indicating difficulty in generating coordinated bimanual motions adapted to the type of box. These results are consistent with the ones obtained in Experiment 1.

VI. CONCLUSIONS

In this paper, we proposed a hierarchical Deep Predictive Learning model composed of three LSTMs for generating motions for coordinated bimanual motions. In our experiments, we compared the hierarchical structure with a single LSTM model. The results showed that our proposed method exhibits higher generalizability than the single LSTM model. We tested our hierarchical model on two different robots, confirming that the approach is effective for bimanual motions without being dependent on hardware. Future works include applying the proposed model to more complex bimanual coordinated tasks, such as furniture assembly or advanced cooking operations, as well as to independent movements of the left and right arms where there is no coordination relationship. Additionally, speeding up the inference process and comparisons with other imitation learning methods, such as Action Chunking with Transformers (ACT) [12] and Diffusion Policy [24], are also important areas to explore.

ACKNOWLEDGMENT

This work was supported by the Comprehensive Research Institute of Science and Engineering at Waseda University, the Super Global University (SGU) Creation Support "Waseda Ocean Initiative," and the JST Moonshot R&D Program JPMJMS2031. We would like to express our gratitude for their support.

REFERENCES

[1] A. Edsinger, C.C. Kemp, "Two Arms Are Better Than One: A Behavior Based Control System for Assistive Bimanual Manipulation," *Lecture Notes in Control and Information Sciences*, vol 370, 2007.
[2] S. Levine, P. Pastor, A. Krizhevsky, D. Quillen, "Learning Hand-Eye Coordination for Robotic Grasping with Deep Learning and Large-Scale Data Collection," *The International Journal of Robotics Research*, Vol.37, No.4-5, pp.421–436, 2018.
[3] 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, W. Zaremba, "Learning dexterous in-hand manipulation," *The International Journal of Robotics Research*, 39–1, pp. 3-20, 2020.
[4] K. Caluwaerts et al., "Barkour: Benchmarking Animal-level Agility with Quadruped Robots," arXiv preprint, arXiv:2305.14654, 2023.
[5] K. Friston: "A theory of cortical responses," *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, vol. 360, 1456, 2005.
[6] P.-C. Yang, K. Sasaki, K. Suzuki, K. Kase, S. Sugano, T. Ogata, "Repeatable Folding Task by Humanoid Robot Worker Using Deep Learning," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 397-403, 2017.
[7] H. Ito, K. Yamamoto, H.Mori, T. Ogata, "Efficient multitask learning with an embodied predictive model for door opening and entry with whole-body control," *Science Robotics*, vol. 7, no. 65, p.eaax8117, 2022.
[8] H. Ichiwara, H. Ito, K. Yamamoto, H. Mori, T. Ogata, "Contact-rich manipulation of a flexible object based on deep predictive learning using vision and tactility," *IEEE International Conference on Robotics and Automation*, pp. 5375-5381, 2011.
[9] K. Suzuki, M. Kanamura, Y. Suga, H. Mori, and T. Ogata, "In-air Knotting of Rope using Dual-Arm Robot based on Deep Learning," Proc. of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp.6724-6731, 2021.
[10] N. Saito, T. Ogata, S. Funabashi, H. Mori and S. Sugano, "How to Select and Use Tools? : Active Perception of Target Objects Using Multimodal Deep Learning," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2517-2524, 2021.
[11] M.A. Diftler et al., "Robonaut 2 - The first humanoid robot in space," 2011 *IEEE International Conference on Robotics and Automation*, pp. 2178-2183, 2011.
[12] F. Zipeng, T.Z. Zhao, F. Chelsea, "Mobile ALOHA: Learning Bimanual Mobile Manipulation with Low-Cost Whole-Body Teleoperation," arXiv preprint, arXiv:2401.02117, 2024.
[13] R.W. Sperry, "Hemisphere deconnection and unity in conscious awareness," *American Psychologist*, 23(10), 723-733, 1968.

[14] L. Liu, Q. Liu, Y. Song, B. Pang, X. Yuan, Q. Xu, "A Collaborative Control Method of Dual-Arm Robots Based on Deep Reinforcement Learning," *Applied Sciences* 11, no. 4: 1816, 2021.
[15] D. Rakita1, B. Mutlu1, M. Gleicher1, L.M. Hiatt, "Shared control–based bimanual robot manipulation," *Science Robotics*, Vol. 4, no. 30, p. eaaw0955, 2019.
[16] S. Kim, K. Ogawa, J. Lv, N. Schweighofer, H. Imamizu, "Neural substrates related to motor memory with multiple timescales in sensorimotor adaptation", *PLOS Biology*, 13(12), 2015.
[17] R.C. Atkinson, R.M. Shiffrin, "Human memory: A proposed system and its control processes," *The psychology of learning and motivation*, 2, 89-195, 1968.
[18] S Hochreiter and J Schmidhuber: "Long Short-Term Memory," *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.
[19] Y. Yamashita and J. Tani, " Emergence of Functional Hierarchy in a Multiple Timescales Recurrent Neural Network Model: A Humanoid Robot Experiment, " *PLoS Computational Biology*, Vol. 4, No. 11, 2008.
[20] F. Krebs and T. Asfour, "A Bimanual Manipulation Taxonomy," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 11031-11038, 2022,
[21] E. Dean-Leon, B. Pierce, F. Bergner, P. Mittendorfer, K. Ramirez-Amaro, W. Burger and G. Cheng, "TOMM: Tactile omnidirectional mobile manipulator," *IEEE International Conference on Robotics and Automation*, pp. 2441-2447, 2017.
[22] G. Cheng, E. Dean-Leon, F. Bergner, J. Rogelio Guadarrama Olvera, Q. Leboutet and P. Mittendorfer, "A Comprehensive Realization of Robot Skin: Sensors, Sensing, Control, and Applications," *Proceedings of the IEEE*, vol. 107, no. 10, pp. 2034-2051, 2019.
[23] K. Kase, K. Suzuki, P.C. Yang, H. Mori, T. Ogata, "Put-in-Box Task Generated from Multiple Discrete Tasks by a Humanoid Robot Using Deep Learning," *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6447-6452, 2018.
[24] Cheng Chi et al., "Diffusion Policy: Visuomotor Policy Learning via Action Diffusion," *The International Journal of Robotics Research*, 2024.