2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Multimodal Time Series Learning of Robots Based on Distributed and Integrated Modalities: Verification with a Simulator and Actual Robots

Hideyuki Ichiwara[1,2], Hiroshi Ito[1,2], Kenjiro Yamamoto[1], Hiroki Mori[2] and Tetsuya Ogata[2]

***Abstract*— We have developed an autonomous robot motion generation model based on distributed and integrated multimodal learning. Since each modality used as a robot's senses, such as image, joint angle, and torque, has a different physical meaning and time characteristic, the generation of autonomous motions using multimodal learning has sometimes failed due to overlearning in one of the modalities. Inspired by the sensory processing of the human brain, our model is based on the processing of each sense performed in the primary somatosensory cortex and the integrated processing of multiple senses in the association cortex and the primary motor cortex. Specifically, the proposed model utilizes two types of recurrent neural networks: sensory RNNs, which learn each sense in a time series, and a union RNN, which communicates with sensory RNNs and learns sensory integration. The simulation results of multiple tasks showed that our model processes multiple modalities appropriately and generates smoother motions with lower jerk than the conventional model. We also demonstrated a chair assembly task by combining fixed motions and autonomous motions with our model.**

I. Introduction

Humans can execute a variety of tasks by utilizing our multiple senses. In robots, as with humans, methods that utilize multiple modalities using deep learning have been proposed to perform tasks. Methods of multimodal learning for robotic manipulation include those using reinforcement learning (RL) [1][2][3] and supervised learning [4][5][6]. The methods using RL require trial and error, and sample efficiency has been an issue, but there are several effective ways to improve sample efficiency [7][8]. However, since trial-and-error is still required, there is a risk of damaging the object in tasks that involve contact with objects. As an alternative, deep predictive learning (DPL) using supervised learning has been attracting attention [4][9][10]. This approach uses human demonstration data to predict multiple modalities in real time and generate autonomous motions by directing the predicted command to the robot. Since DPL does not require trial-and-error, the risk of damage to the robot or object is low.

There are various senses that a robot can use, such as vision (RGB images), kinesthesia (joint angles and torque), tactility (tactile sensor values), and hearing (audio). Numerous tasks have been achieved by combining these. However, learning multiple modalities is not easy, as each modality has a different physical meaning and time characteristics. There are several techniques to efficiently learn each modality, and the common practice is to process each modal using these techniques and to integrate the processed modals in a single layer for learning [9][11][12][13]. There is also a method for integrating multiple modalities that focuses on differences in the modal time constants and uses multiple timescale recurrent neural networks (MTRNNs) [14], which can have multiple time constants [15][16]. However, the time constants need to be set in advance, which is not easy when many different modals are used.

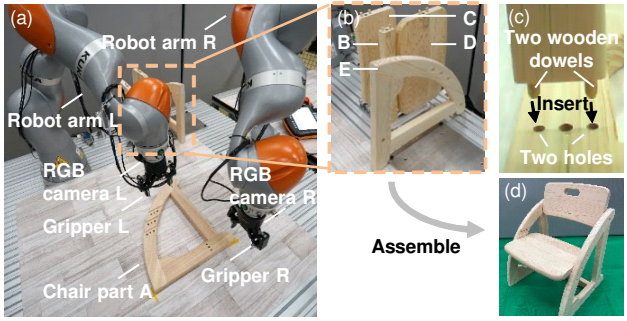

Fig. 1. Demonstration setup of a chair assembly task with actual robots. (a) Overview. (b) Chair parts. (c) Wooden dowels and holes. Two wooden dowels are inserted into two holes to attach the parts. (d) Finished product.

Inspired by the sensory processing of the human brain, we focus on the processing of each modality performed in the primary somatosensory cortex and the integration processing of multiple modalities performed in the association cortex and the primary motor cortex [17]. Specifically, we developed a motion generation model that uses two types of recurrent neural network (RNN): sensory RNNs (SRNNs), which learn each sense in a time series, and a union RNN (URNN), which communicates with SRNNs and learns sensory integration. In this study, we used a simulator to compare our model with a conventional one in several tasks involving contact with objects. We also demonstrated the task of assembling a chair in the environment shown in Fig. 1.

The main contributions of this research are as follows. 1) We developed an autonomous motion generation model with multimodal learning based on distributed and integrated modalities. 2) The effectiveness of our model was demonstrated by its high success ($\geq$ 83.3%) rates and smooth motion in multiple tasks using a simulator. 3) We demonstrated a chair assembly task using two real robots by combining fixed motions and autonomous motions by our model.

[1]Hideyuki Ichiwara, Hiroshi Ito and Kenjiro Yamamoto are with Research & Development Group, Hitachi, Ltd., Ibaraki, 312-0034, Japan hideyuki.ichiwara.bn@hitachi.com

[2]Hideyuki Ichiwara, Hiroshi Ito, Hiroki Mori and Tetsuya Ogata are with Department of Intermedia Art and Science School of Fundamental Science and Engineering, Waseda University, Tokyo, 169-855, Japan ogata@waseda.jp

II. Related Work

A. Processing methods for each modality

The modalities considered here include vision, tactility, kinesthesia, and hearing, and there are various processing methods that consider the features of each. For vision, there is a visual attention model that extracts location information from camera images, which is important for tasks [1][18]. For tactility, CNN-based processing based on two-dimensional feature information (similar to images) has been proposed [9][12]. For kinesthesia, softmax transformation has been utilized to enhance learning by quantizing motions represented by continuous values to a sparse representation [19]. An implicit kinetic policy, which focuses on the difference between the properties of joint space and Cartesian space, has also been proposed [20]. For hearing, there is a method that converts noisy, high-dimensional raw waveform features into spectrograms delimited by a fixed window width and processed by 1D-CNN [11].

B. Integrating multiple modalities

There are several studies on motion learning that integrate multiple modalities. Examples include the use of force and kinesthesia for drawing tasks [5] and rubbing tasks [16]. Other work has added vision to force and kinesthesia for unzipping a fabric bag [9], wiping 3D objects [21], and manipulating clothing [22]. In many of these studies, each modality processed in a previous stage is integrated in a feedforward manner with a single NN layer to generate motions. Ito et al. [16] focused on the difference in time characteristics between force and kinesthesia and came up with a method that uses MTRNN with three recurrent layers featuring different time constants. In their method, force sense is considered to have a fast time constant and is input to the fast layer, while motion sense is considered to have a relatively slow time constant and is input to the middle layer. However, it is difficult to decide on the time constant and where to place the modality in the layer when the number of modalities increases. Saito et al. [15] proposed a method in which force, touch, and motion are input to the fast time constant layer and only the initial image is input to the slow time constant layer when wiping 3D objects with occlusion. However, because only the initial image is used, it is difficult to cope with visual changes, and there are issues similar to those facing Ito et al.'s method. In short, there is currently no unified methodology to integrate each processed modality.

III. Method

In the proposed method, we utilize DPL to mitigate the risk of component breakdown. DPL uses a neural network to predict multiple modalities at the next time $t+1$ that include control commands of the robot from those at the current time t. The predicted commands are then sequentially sent to the robot to generate the motion. The model is trained using the time-series data of the actions to be learned by the robot (e.g., human operation) as training data.

Figure 2(a) shows the DPL model (visual attention model [9][18]) used in this study. The inputs are an image $\boldsymbol{v}_t$ at time t, control commands to the robot $\boldsymbol{c}_t$, and other modalities $\boldsymbol{m}_t$. The output is each predicted modal information $\hat{\boldsymbol{f}}^p_{t+1}, \hat{\boldsymbol{c}}_{t+1}, \hat{\boldsymbol{m}}_{t+1}$ at time $t+1$. Note that the hat indicates the output of the neural network, and the absence of the hat indicates the true value (measured value). The model consists of three parts: an encoder, a recurrent part, and a decoder. The encoder extracts position coordinates in the image as image features from the camera image, the recurrent part integrates each modal and learns time-series changes, and the decoder predicts the image. The encoder utilizes CNN and soft argmax [23] to output the position coordinates $\hat{\boldsymbol{f}}^c_t$ (called attention points in this paper) of points in the image from the input image $\boldsymbol{v}_t$. On another route, feature maps are obtained using CNN and input to the decoder. Next, in the recurrent part using fully connected neural networks (FCNNs) and an RNN, the coordinates, joint angles, and other modal information of the point of attention at time t are used to predict the attention points, joint angles, and other modal information at the next time $t+1$. The decoder predicts the image at the next time $t+1$ based on the predicted attention points and the feature maps obtained from the encoder part. Specifically, heat maps with high intensity at the predicted attention points are generated, and the image at the next time $t+1$ is predicted from the feature map weighted by these maps. This is to support image prediction by using image features in the vicinity of the attention points. In our method, we utilize a new recurrent part that integrates multiple modalities.

A. Conventional recurrent part

Figure 2(b)(i) shows how the conventional recurrent part integrates multiple modalities using an RNN and predicts each modality using the FCNNs. The term RNN here refers to as a basic network capable of learning time series. Gated recurrent units [24], long-short term memory (LSTM) [25], etc. can be used; in this study, we opted for LSTM. In the RNN, the hidden state $\boldsymbol{h}_t$ is computed using the modality $\hat{\boldsymbol{f}}^c_t, \boldsymbol{c}_t, \boldsymbol{m}_t$ at time t and the hidden state one time ago $\boldsymbol{h}_{t-1}$ as follows.

$$\boldsymbol{h}_t = \mathrm{RNN}(\hat{\boldsymbol{f}}^c_t, \boldsymbol{c}_t, \boldsymbol{m}_t, \boldsymbol{h}_{t-1}). \tag{1}$$

Each predicted modal information $\hat{\boldsymbol{f}}^p_{t+1}, \hat{\boldsymbol{c}}_{t+1}, \hat{\boldsymbol{m}}_{t+1}$ at the next time $t+1$ is then predicted using a hidden state $\boldsymbol{h}_t$ that contains the multimodal information and an FCNN for each modal. In this way, multimodal information such as visual information, force information, and joint angles can be integrated to generate diverse motions by predicting joint angles at the next time. However, it is not easy to integrate and process multiple modal information with a single RNN, and sometimes it does not work well due to overlearning in one of the modalities.

B. Proposed recurrent part

The new part we propose consists of FCNNs and two types of RNNs: an SRNN, which learns each modality in time series, and a URNN, which communicates with SRNNs and learns sensory integration. In the URNN, each abstracted

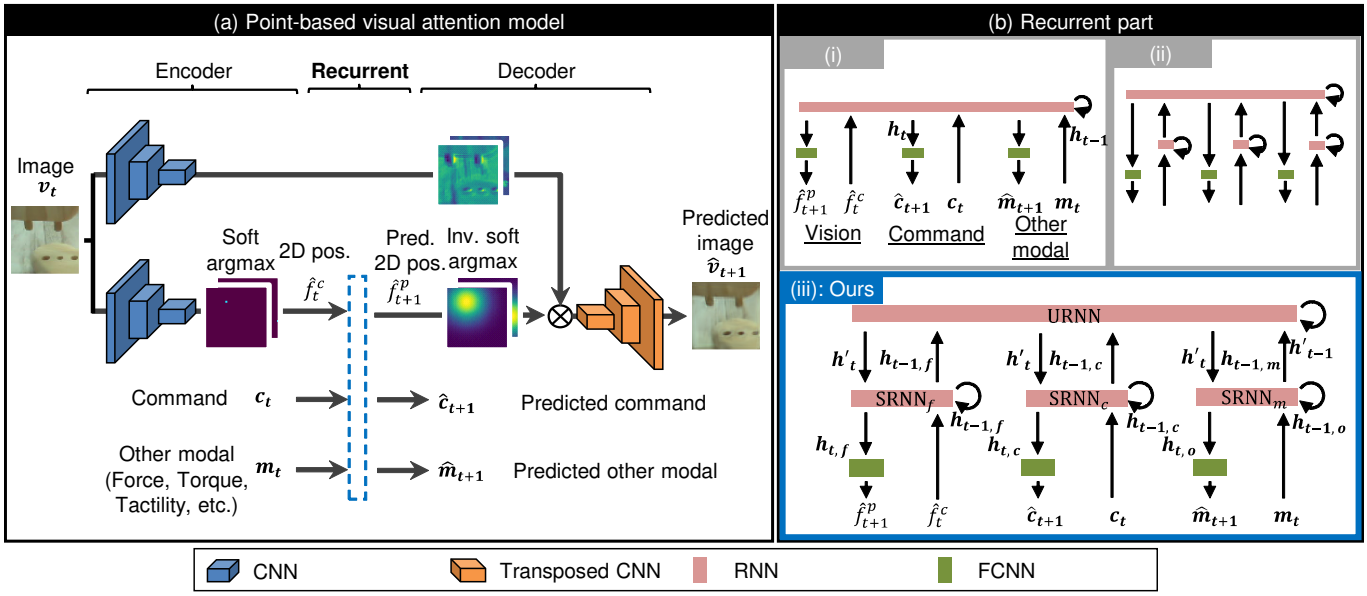

Fig. 2. Proposed DPL model for generating motion. (a) Point based visual attention model. (b) Recurrent part. (i) Conventional. Integration of all modalities in a single RNN. (ii) Separate RNNs for each modal and for integration. (iii) Our model. Separate RNNs for each modal (SRNN) and for integration (URNN), and feed back from URNN to SRNN.

modality, a hidden state $\boldsymbol{h}_{t-1,i}(i = f, c, m)$, is integrated and processed, and the hidden state at time t passed to each SRNN is updated as follows.

$$\boldsymbol{h}'_t = \mathrm{URNN}(\boldsymbol{h}_{t-1,i}, \boldsymbol{h}'_{t-1}) \quad (i = f, c, m). \tag{2}$$

In each SRNN, the hidden state $\boldsymbol{h}'_t$ is calculated using the hidden state $\boldsymbol{h}'_{t,i}(i = f, c, m)$ updated by the URNN that contains multiple modalities.

$$\boldsymbol{h}_{t,f} = \mathrm{SRNN}_f(\hat{\boldsymbol{f}}^c_t, \boldsymbol{h}_{t-1,f}, \boldsymbol{h}'_t), \tag{3}$$
$$\boldsymbol{h}_{t,c} = \mathrm{SRNN}_c(\boldsymbol{c}_t, \boldsymbol{h}_{t-1,c}, \boldsymbol{h}'_t), \tag{4}$$
$$\boldsymbol{h}_{t,m} = \mathrm{SRNN}_m(\boldsymbol{m}_t, \boldsymbol{h}_{t-1,m}, \boldsymbol{h}'_t). \tag{5}$$

Each predicted modal information at the next time $t+1$ is then predicted using the hidden state $\boldsymbol{h}_{t,i}$ and the FCNN for each modal.

Since each SRNN has an inner loop for each modal, the time-varying characteristics of each can be considered. Furthermore, because integrated multimodal information is obtained from the URNN, it is possible to predict each modal with other modal information. This is akin to the primary somatosensory cortex, which processes each sense and is in contact with the parietal association cortex and the motor cortex [17]. It is also consistent with the fact that the motor cortex, which integrates sensations and outputs motor commands, sends movement-related information to the primary somatosensory cortex [26].

Note that this is different from the model shown in Fig. 2(b)(ii), where RNNs are prepared for each modal and integrated in a feedforward manner. The proposed model in Fig. 2(b)(iii) has feedback from the URNN to the SRNN, while Fig. 2(b)(ii) does not. Figure 2(b)(ii) splits the RNN, but in the end, the RNN in the latter stage outputs the prediction for each modal, making it difficult to take advantage of the individual inner loops or to take the time characteristics into account. Therefore, slow time-varying modals will presumably be affected by fast time-varying modals, thus resulting in noisy predictions.

C. Loss function

The loss function in our method is defined as follows:

$$g = \sum_{t \in T-1} (g_{t,v} + g_{t,c} + g_{t,m} + g_{t,f}), \tag{6}$$
$$g_v = \frac{1}{H \times W \times C} \|\hat{\boldsymbol{v}}_{t+1} - \boldsymbol{v}_{t+1}\|_2^2, \tag{7}$$
$$g_c = \frac{1}{M} \|\hat{\boldsymbol{c}}_{t+1} - \boldsymbol{c}_{t+1}\|_2^2, \tag{8}$$
$$g_m = \frac{1}{D} \|\hat{\boldsymbol{m}}_{t+1} - \boldsymbol{m}_{t+1}\|_2^2, \tag{9}$$
$$g_f = \frac{1}{K} \|\hat{\boldsymbol{f}}^p_{t+1} - \hat{\boldsymbol{f}}^c_{t+1}\|_2^2, \tag{10}$$

where the sequence length of the training data is T, the time is t, the image is $\boldsymbol{v} \in \mathbb{R}^{H \times W \times C}$, the joint angle $\boldsymbol{c} \in \mathbb{R}^M$ and the coordinates of the attention points $\hat{\boldsymbol{f}} \in \mathbb{R}^K$. g_v, g_c, g_m are prediction errors of the image, the command, and the other modality using the mean square error, respectively. The g_f is an auxiliary loss function, which is added so that the attention points output by the recurrent part are obtained on the basis of the prediction of attention points. Note that since the training is end-to-end, it is not necessary to teach the attention points, and g_f is calculated between the outputs of the neural network.

IV. Experiments

A. Experimental setup

We used NVIDIA's Isaac Sim [27], which was chosen because of its photorealistic nature; if it was not photoreal-

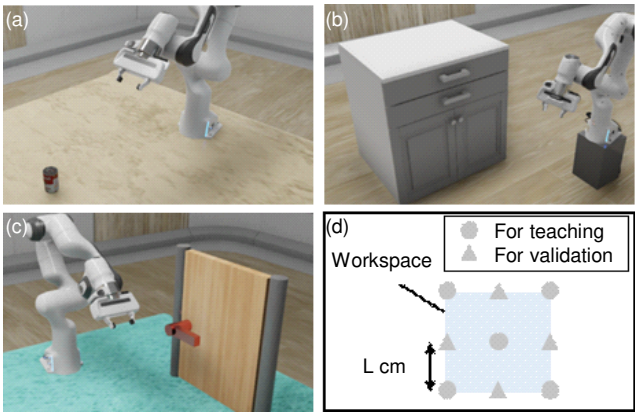

Fig. 3. Experimental setup. (a) Picking. (b) Drawer opening. (c) Door opening. (d) Object position for teaching and verification.

TABLE I

NETWORK PARAMETERS.

	Layer type	Parameter	Output shape
Encoder	CNN 1_1	k=3, s=1, c=16	62×62×16
	CNN 1_2	k=3, s=1, c=32	60×60×16
	CNN 1_3	k=3, s=1, c=K	58×58×K
	Soft argmax	-	K×2
	CNN 2_1	k=3, s=1, c=16	62×62×16
	CNN 2_2	k=3, s=1, c=32	60×60×16
	CNN 2_3	k=3, s=1, c=K	58×58×K
Recurrent	SRNN	n=50	50
	FCNN*	n=2×K, 8, 8, 8	K, 8, 8, 8
	URNN	n=20	20
Decoder	Inv soft argmax	-	58×58×K
	Transposed CNN 1	k=3, s=1, c=32	60×60×32
	Transposed CNN 2	k=3, s=1, c=16	62×62×16
	Transposed CNN 3	k=3, s=1, c=3	64×64×3

c, k, s, and n denote number of channels, kernel, strides and output dimensions (nodes), respectively
* There are four FCNNs (one for each modality).

istic, the images would be too easy to predict and the trends would not match those of the real environment. The robot was Panda by Franka Emika. The camera was placed in a position where it could see the workspace. We used RGB color images with a resolution of 64×64, so $H = 64, W = 64, C = 3$ in Eq. (7). Four modalities were measured and input to the model: camera image, measured joint angle, commanded joint angle, and joint torque. The robot arm has seven axes and the end-effector is a two-fingered gripper, so $M = 8$ in Eq. (8). The number of attention points varied depending on the task. Figure 3 shows the experimental setup of the simulator. Three tasks that involve contact with the object were selected: (a) Picking, (b) Drawer opening, and (c) Door opening. We compared the three models (i), (ii), (iii) and tested each task and model 36 times at random object locations in the workspace (shown in (d)). $L = 8, 8, 5$ [cm] in (a), (b), and (c), respectively.

The training data consisted of 24 demonstration data: four for each of the points for teaching and one for each of the points for validation. The training data were acquired by a human operating the robot using a HTC VIVE VR controller. No pre-training was performed for any of the models, and

TABLE II

TASK SUCCESS RATE FOR EACH TASK AND MODEL IN EXPERIMENT 1.

Model \ Task	Picking	Drawer opening	Door opening
(i)	83.3%	13.8%	86.7%
(ii)	55.5%	61.1%	86.1%
(iii): Ours	**88.8%**	**83.3%**	**100.0%**

TABLE III

AVERAGE SUM OF SQUARES OF JERK FOR EACH TASK AND MODEL IN EXPERIMENT 1. UNIT IS $[\mathrm{m/s^3}]$.

Model \ Task	Picking	Drawer opening	Door opening
(i)	70.3±6.4	242.0±23.2	79.7±2.1
(ii)	66.9±5.3	194.3±4.3	78.9±3.2
(iii): Ours	65.1±5.3	**161.9±4.8**	**71.7±3.2**

all of them were trained in an end-to-end fashion. Table I shows the network parameters of our model (iii), where K is the number of attention points and depends on the tasks. The network parameters of model (i) (ii) were alomost the same as Table I. The number of nodes in the RNN of (i) was set to 120 so that the number of weights in the recurrent part was the same. We set $K = 4, 8, 4$ for picking, drawer opening, and door opening, respectively.

B. Results

Table II lists the success rate for each task and model. For all tasks, our model (iii) had the highest success rate. In the conventional model (i) and in model (ii), there were several cases of failure because of the hand tip slipping against the object or unstable movement as a result of unstable attention.

Table III shows the average sum of squares of the jerk for each task and model [28], as an index of the smoothness of the motion. Note that the computation period of the neural network was 10 Hz and the commands were sent at this period, but the calculation of the jerk refers to the data obtained at the measurement period of 50 Hz. The jerk of our model was significantly lower in the two tasks, indicating that it generated motion more smoothly than the other models. Since the demonstration data are manipulated by human, the jerk is considered to be low. This finding suggests that our model is able to generate motions more in line with the demonstration data by appropriately processing each modality. On the other hand, the jerk could be explicitly included in the loss function. However, there are two problems: the time resolution is coarse because the jerk cannot be computed with a sampling period shorter than the sampling period of the training data, and the balance with other loss functions must be adjusted. It is important to note that our model can generate smooth motion with smaller jerks as a result of predictive learning that does not include jerks in the loss function.

As for the variability of attention, Table IV shows the rate of time-varying attention points for each task model. The time-varying rate of the attention points with our model was

TABLE IV

TIME-VARYING RATE OF ATTENTION POINTS FOR EACH TASK AND MODEL IN EXPERIMENT 1. UNIT IS $[-/\mathrm{s}]$.

Model \ Task	Picking	Drawer opening	Door opening
(i)	0.60±0.01	1.79±0.13	0.95±0.05
(ii)	0.60±0.01	1.15±0.81	0.80±0.05
(iii): Ours	**0.40±0.05**	**0.71±0.05**	0.80±0.05

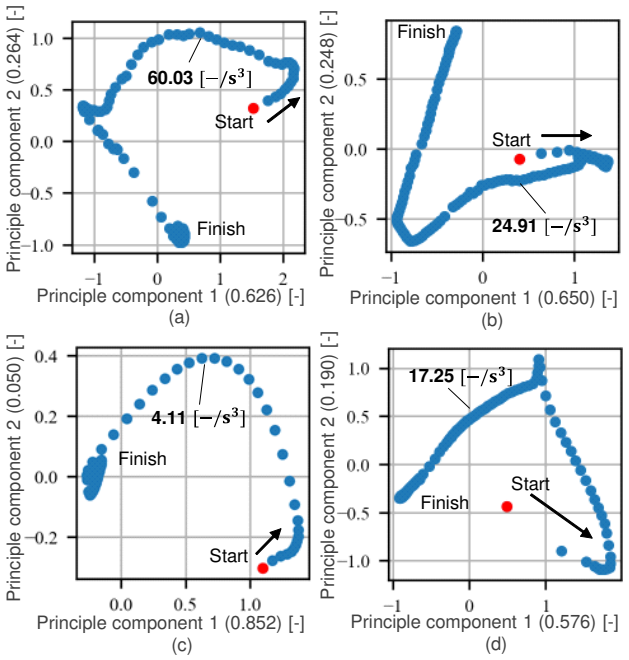

Fig. 4. PCA results for hidden states of RNN at the drawer opening task. (a) About $\boldsymbol{h}_t$ of model (i). (b) About $\boldsymbol{h}_{t,v}$ of our model (iii). (c) About $\boldsymbol{h}_{t,c}$ of our model (iii). (d) About $\boldsymbol{h}'_t$ of our model (iii).

significantly lower in the two tasks, indicating stable attention. Although all models had the same visual processing system, the proposed model yielded more stable attention points. This suggests that our model configuration contributes to the appropriate modal processing.

To examine whether the time characteristics of each modal can be considered, we confirmed the hidden state of the recurrent part. Figure 4 shows the results of principal component analysis (PCA) for the hidden state of the inner loop of the RNN at the drawer opening task. (a) is the PCA result about $\boldsymbol{h}_t$ of model (i). A qualitatively non-smooth representation is obtained, which is expected to be affected by fast time-varying modals. Quantitatively, the sum of jerk was 60.03 $[-/\mathrm{s}^3]$. (b)(c)(d) are the PCA result about $\boldsymbol{h}_{t,v}, \boldsymbol{h}_{t,c}$ and $\boldsymbol{h}'_t$. The sum of jerk of (b)(c)(d) are 24.91, 4.11, and 17.25$[-/\mathrm{s}^3]$, each with different time characteristics. In particular, even if the joint motions are continuous, the change of attention in vision may become discontinuous when the position to which attention should be directed changes because of switching of motions. Therefore, it is considered that the change in the hidden state of (b) regarding

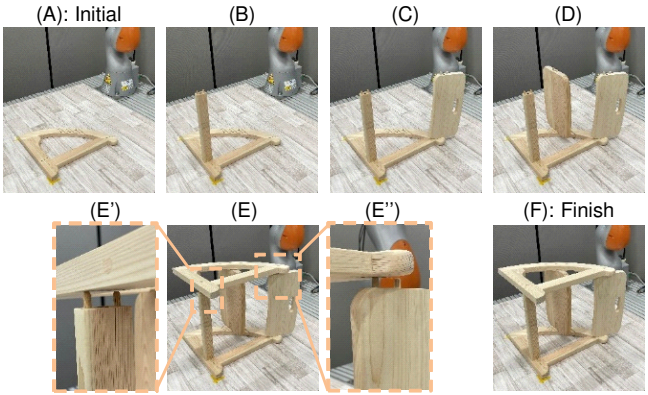

Fig. 5. Flow of chair assembly. The initial state is where part A is placed, and parts B, C, D, and E are then attached in order.

the attention points is not relatively smooth because the time characteristics of the joint angle and the attention point are different. In this way, the time characteristics are different for each modal, but each SRNN can take into consideration the appropriate characteristics. In addition, URNN in (d) has fast-time characteristics, and it is thought that the information is transmitted to the SRNN to the extent that it does not impair fast-time characteristics of vision and does not interfere with other modal predictions.

V. DEMONSTRATION

A. Setup

As shown in Fig.1, the setup consisted of two KUKA LBR iiwa 14 R820 robot arms. An Robotiq 2F-85 Adaptive Gripper was attached to each arm. RGB camera Buffalo's BSW500M Series and realsense L515 were placed on the gripper, respectively. The chair was a children's chair manufactured by IKONIH and consisted of five wooden parts. Although a jig was created to stand the parts and place them in a position where they could be grasped by the arm, these part-specific jigs are not used to strictly fix the position with generality in mind. Figure 5 shows the assembly flow. The initial state is where part A is placed, and then parts B, C, and D are attached in order. Part E is then inserted into D. After this, B and C are not inserted into E, but since D is inserted, we found that it can be inserted by applying a slight perturbation. Therefore, we created a motion to apply perturbation by grabbing roughly in advance, and then executed it for the insertion. Finally, by pushing E from the top, all wooden dowels are completed inserted all the way to the back. The overall task design policy is to combine fixed motions (mainly point-to-point) with autonomous motions generated by our model.

We chose this task because inserting the wooden dowels is difficult and requires the use of learning behaviors. It is difficult because postural deviations in addition to the position have an impact on the success or failure of the task, because the position is not strictly fixed by the jig and there is a deviation during grasping, and because part A moves during insertion, causing the position deviation to

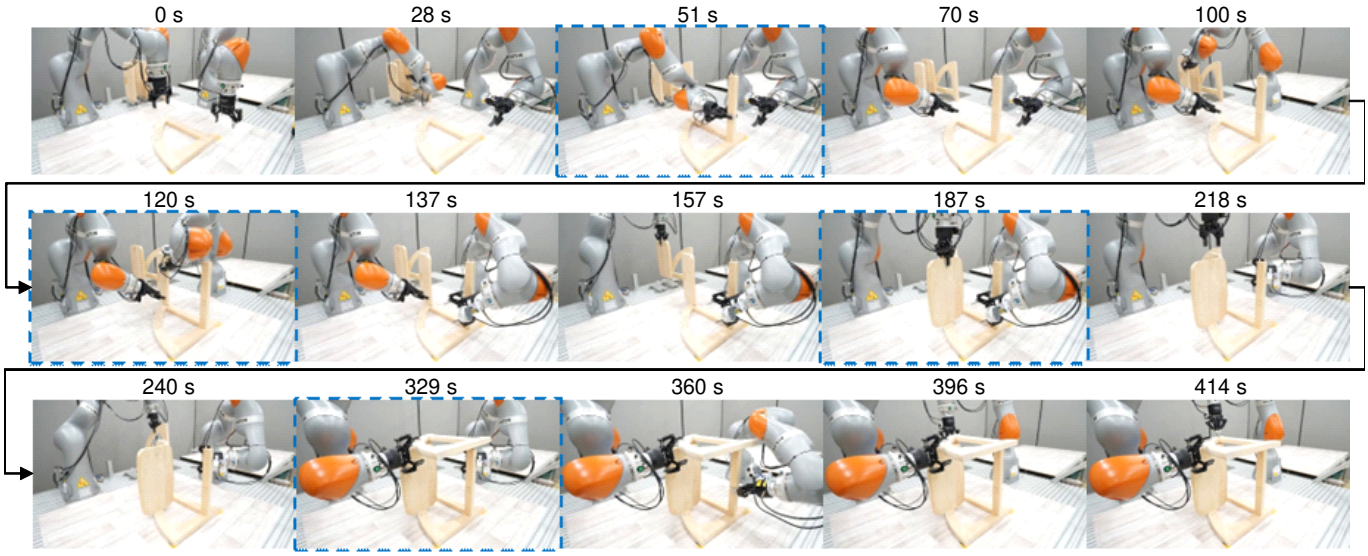

Fig. 6. Snapshots of the chair assembly task.

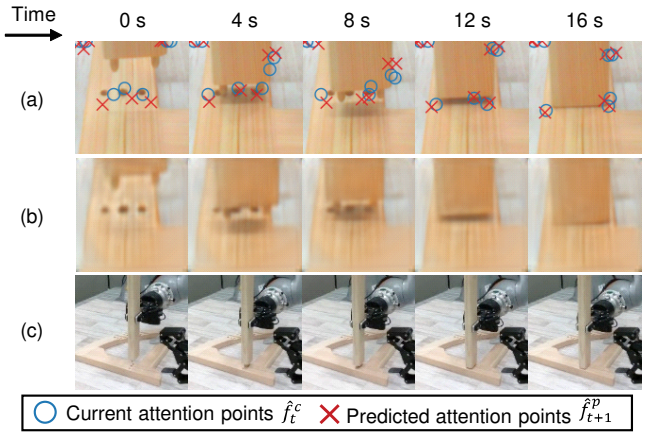

Fig. 7. Snapshots of part B in the insertion task. (a) Input image and attention points. (b) Predicted image. (c) Birds-eye view (not for control).

expand as the task progresses. This insertion task requires proper processing of multiple modalities and is well suited to demonstrate the effectiveness of the proposed model. Compared to the usual peg-in-hole task, where the holes and pegs have a one-to-one correspondence, this task is more difficult due to severe postural constraints and the requirements for processing kinesthesia. In addition, there are occlusions of the wooden dowels and holes during the task, and shadows are generated depending on the position of the parts, so the vision changes dynamically, requiring the appropriate processing of vision. Furthermore, there is contact between the wooden dowel and the hole, so appropriate processing of force is also necessary.

The four insertion tasks (B), (C), (D), and (E) are learned indivisually. Thirty-six items of data were collected for each part and used as training data. The data to be measured and input to the model consisted of four types: camera images, measured joint angles, commanded joint angles, and 6-axis forces received by the end-effector. The model parameters were the same as in the previous experiment.

B. Results

Figure 6 shows the snapshots of the task. The pictures indicated by the blue frame are due to the learning operation. Since the individual parts are not fixed, they may shift during operation, but this is compensated by the learning operation, and a series of operations could be achieved.

The success rate of the insertion task for part B was high, 91.7% (110/120 times). The evaluation of the other parts is omitted here because they are similar tasks. Figure 7 shows (a) the input images and the attention points, (b) the predicted images, and (c) the bird's eye view during the task. This is the case where part B is on the camera side rather than the hole, and occlusion occurs. The attention points are on the hole and the part, which are important for the task. During the task, the hole occludes, and a shadow is generated by the approach of the part, but the operation does not become unstable, and the task is successful.

VI. Conclusion

Inspired by the sensory processing of the human brain, we developed a DPL model that refers to the distributed and integrated processing of multiple modalities. Simulation results demonstrated the effectiveness of the model in terms of success rate and smoothness of motion in multiple tasks. We also demonstrated a successful chair assembly task, which included occlusion and shadow generation, with two real robots.

Acknowledgement

Computational resource of AI Bridging Cloud Infrastructure (ABCI) provided by National Institute of Advanced Industrial Science and Technology (AIST) was used.

REFERENCES

[1] C. Finn, X. Y. Tan, Y. Duan, T. Darrell, S. Levine, and P. Abbeel, "Deep spatial autoencoders for visuomotor learning," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 512–519.

[2] Y. Tsurumine, Y. Cui, E. Uchibe, and T. Matsubara, "Deep reinforcement learning with smooth policy update: Application to robotic cloth manipulation," *Robotics and Autonomous Systems*, vol. 112, pp. 72–83, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0921889018303245

[3] H. van Hoof, N. Chen, M. Karl, P. van der Smagt, and J. Peters, "Stable reinforcement learning with autoencoders for tactile and visual data," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016, pp. 3928–3934.

[4] K. Noda, H. Arie, Y. Suga, and T. Ogata, "Multimodal integration learning of robot behavior using deep neural networks," *Robotics and Autonomous Systems*, vol. 62, no. 6, pp. 721–736, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0921889014000396

[5] T. Adachi, K. Fujimoto, S. Sakaino, and T. Tsuji, "Imitation learning for object manipulation based on position/force information using bilateral control," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 3648–3653.

[6] H. Kim, Y. Ohmura, and Y. Kuniyoshi, "Transformer-based deep imitation learning for dual-arm robot manipulation," *CoRR*, vol. abs/2108.00385, 2021. [Online]. Available: https://arxiv.org/abs/2108.00385

[7] R. Shah and V. Kumar, "Rrl: Resnet as representation for reinforcement learning," *arXiv preprint arXiv:2107.03380*, 2021.

[8] P. Wu, A. Escontrela, D. Hafner, K. Goldberg, and P. Abbeel, "Daydreamer: World models for physical robot learning," *arXiv preprint arXiv:2206.14176*, 2022.

[9] H. Ichiwara, H. Ito, K. Yamamoto, H. Mori, and T. Ogata, "Contact-rich manipulation of a flexible object based on deep predictive learning using vision and tactility," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 5375–5381.

[10] H. Ito, K. Yamamoto, H. Mori, and 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. eaax8177, 2022. [Online]. Available: https://www.science.org/doi/abs/10.1126/scirobotics.aax8177

[11] M. Du, O. Y. Lee, S. Nair, and C. Finn, "Play it by ear: Learning skills amidst occlusion through audio-visual imitation learning," 2022. [Online]. Available: https://arxiv.org/abs/2205.14850

[12] J. Hansen, F. Hogan, D. Rivkin, D. Meger, M. Jenkin, and G. Dudek, "Visuotactile-rl: Learning multimodal manipulation policies with deep reinforcement learning," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 8298–8304.

[13] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *CoRR*, vol. abs/1504.00702, 2015. [Online]. Available: http://arxiv.org/abs/1504.00702

[14] Y. Yamashita and J. Tani, "Emergence of functional hierarchy in a multiple timescale neural network model: a humanoid robot experiment," *PLoS computational biology*, vol. 4, no. 11, p. e1000220, 2008.

[15] N. Saito, T. Shimizu, T. Ogata, and S. Sugano, "Utilization of image/force/tactile sensor data for object-shape-oriented manipulation: Wiping objects with turning back motions and occlusion," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 968–975, 2022.

[16] H. Ito, T. Kurata, and T. Ogata, "Sensory-motor learning for simultaneous control of motion and force: Generating rubbing motion against uneven object," in *2022 IEEE/SICE International Symposium on System Integration (SII)*, 2022, pp. 408–415.

[17] A. Starr and L. G. Cohen, " ' gating ' of somatosensory evoked potentials begins before the onset of voluntary movement in man," *Brain research*, vol. 348, no. 1, pp. 183–186, 1985.

[18] H. Ichiwara, H. Ito, K. Yamamoto, H. Mori, and T. Ogata, "Spatial attention point network for deep-learning-based robust autonomous robot motion generation," *arXiv preprint arXiv:2103.01598*, 2021.

[19] J. Hwang and J. Tani, "Seamless integration and coordination of cognitive skills in humanoid robots: A deep learning approach," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 2, pp. 345–358, 2018.

[20] A. Ganapathi, P. Florence, J. Varley, K. Burns, K. Goldberg, and A. Zeng, "Implicit kinematic policies: Unifying joint and cartesian action spaces in end-to-end robot learning," *arXiv preprint arXiv:2203.01983*, 2022.

[21] N. Saito, D. Wang, T. Ogata, H. Mori, and S. Sugano, "Wiping 3d-objects using deep learning model based on image/force/joint information," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 10 152–10 157.

[22] K. Kawaharazuka, A. Miki, M. Bando, K. Okada, and M. Inaba, "Dynamic cloth manipulation considering variable stiffness and material change using deep predictive model with parametric bias," *Frontiers in Neurorobotics*, vol. 16, 2022.

[23] D. C. Luvizon, D. Picard, and H. Tabia, "2d/3d pose estimation and action recognition using multitask deep learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5137–5146.

[24] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[26] T. Umeda, T. Isa, and Y. Nishimura, "The somatosensory cortex receives information about motor output," *Science Advances*, vol. 5, no. 7, p. eaaw5388, 2019. [Online]. Available: https://www.science.org/doi/abs/10.1126/sciadv.aaw5388

[27] NVIDIA. (2022) Nvidia isaac sim — nvidia developer. [Online]. Available: https://developer.nvidia.com/isaac-sim

[28] T. Flash and N. Hogan, "The coordination of arm movements: an experimentally confirmed mathematical model," *Journal of Neuroscience*, vol. 5, no. 7, pp. 1688–1703, 1985. [Online]. Available: https://www.jneurosci.org/content/5/7/1688