IEEE Robotics and Automation Letters (RAL) paper presented at
2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Uncertainty-Aware Haptic Shared Control With Humanoid Robots for Flexible Object Manipulation

Takumi Hara, Takashi Sato, *Senior Member, IEEE*, Tetsuya Ogata, *Member, IEEE*, and Hiromitsu Awano, *Associate Member, IEEE*

***Abstract*—We propose a haptic shared control system that predicts human manipulation intentions using a neural network and adaptively presents haptic guidance to achieve smooth robot control remotely. Although the haptic shared control has garnered increasing attention as a method to improve operability in remote operations, incorrect guidance can worsen operability. In this study, we dynamically switch the strength of haptic guidance presentation depending on the uncertainty of the inference results of the neural network. Thus, we weaken the haptic guidance presentation strength for predictions in which the neural network lacks confidence and strengthen it for those with high confidence, thereby achieving guidance presentation that does not impede human manipulation. As a result of experiments using the Nextage OPEN upper-body humanoid robot, in a task involving folding a flexible object, we succeeded in reducing task execution time by 17.1% compared to that with an existing method that determines the strength of haptic guidance presentation without considering the confidence of the neural network.**

***Index Terms*—Imitation learning, haptics and haptic interfaces, human-centered automation.**

I. Introduction

BECAUSE of the worldwide decline in birth rates and the trend towards an aging society, the shortage of labor force is increasingly concerning. In addition, with the recent pandemic experience, there has been a widespread diversification of work styles such as telework. These circumstances have led to the continued research and development of automated control robots to improve labor productivity and realize a ubiquitous society.

Conventional robots have mainly focused on manipulating solid objects, and their activity fields have been centered on production sites, such as factories. However, because of the increase in caregiving and household burden caused by the declining birth rate, aging society, and trend towards nuclear families, the practical use of service robots to address these issues is highly desired. One of the barriers to realizing service robots is the presence of flexible materials such as paper and cloth in the workspace. Generally, the manipulattion of such materials is difficult for robots because their shape is affected by the end-effector trajectory of the robot during manipulation. To address this issue, a method of controlling robots using neural networks that learn human-demonstrated actions through end-to-end (E2E) learning has been reported [1], [2], [3].

However, even after using E2E learning for fully autonomous robots, the risks of inference errors and associated malfunctions still exist. For example, in the field of caregiving, robot malfunctions pose a potential direct threat to humans, and thus, the development of remotely operated robots controlled by humans is active in such fields [4]. However, even with the development of control interfaces, the remote operation of multi-DoF robots still imposes a significant burden on operators. One factor is the incomplete transmission of human sensation from the robot, as it is challenging to perceive depth when operating through a monitor [5]. Another reason is that the joint structure of the robot arm is considerably different from that of the human arm, making it difficult for the operator to infer the movements of the robot arm in a remote environment [6], [7]. Therefore, to reduce the burden on the operator, shared control technology has been devised, which enables humans and AI to share the control of the robot, and humans to intervene in the operation when the autonomous control of the robot by AI is difficult.

Based on the method of interaction between human and AI, shared control can be categorized into two types, namely haptic shared control and mixed-input shared control [8]. Haptic shared control is an operational format in which a human and an AI share the same operation interface to control a machine, and the human determines the operation input while continuously receiving haptic feedback generated by AIs. An example of this type of operation is the function that prevents a car from deviating from the lane by allowing the car to perform steering operations. Conversely, mixed-input shared control is an operational format in which the human and AI do not share the same operation interface; however, their operation inputs are combined in the control device. The collision avoidance braking function in an automobile falls under this category.

Manuscript received 12 April 2023; accepted 6 August 2023. Date of publication 18 August 2023; date of current version 29 August 2023. This letter was recommended for publication by Associate Editor M. Selvaggio and Editor J.-H. Ryu upon evaluation of the reviewers' comments. This work was supported by JSPS KAKENHI under Grant 21H03409. *(Corresponding author: Takumi Hara.)*

Takumi Hara is with the Depertment of Informatics, Graduate School of Informatics, Kyoto University, Sakyo, Kyoto 606-8501, Japan (e-mail: thara@easter.kuee.kyoto-u.ac.jp).

Takashi Sato and Hiromitsu Awano are with the Department of Communication and Computer Engineering, Graduate School of Infomatics, Kyoto University, Sakyo, Kyoto 606-8501, Japan (e-mail: takashi@i.kyoto-u.ac.jp; awano@i.kyoto-u.ac.jp).

Tetsuya Ogata is with the Department of Intermedia Art and Science, School of Fundamental Science and Engineering, Waseda University, Shinjuku, Tokyo 169-8555, Japan (e-mail: ogata@waseda.jp).

This letter has supplementary downloadable material available at https://doi.org/10.1109/LRA.2023.3306668, provided by the authors.

Digital Object Identifier 10.1109/LRA.2023.3306668

In response to the increasing utilization of shared control in the automotive field, shared control is being applied to arm-type robots that require more complex control [9]. In this study, solid object grasping by a remote-controlled robot is achieved by presenting haptic guidance, such as moving a robot gripper to a grasping point candidate based on inferences made by AI, to a remote operator through a haptic feedback device. However, this approach has two issues. One issue is that the uncertainty of AI inferences when presenting haptic guidance to operators is not considered, which may worsen the operability by presenting incorrect haptic guidance. Another issue is that it cannot handle objects, such as clothing whose shape changes considerably during manipulation. Grasping point candidates are generated using still images of the entire work area before the robot moves, and the objects whose shape changes considerably during robot operation cannot be handled.

This letter proposes a haptic-shared control system for an upper-body humanoid robot based on imitation learning. The proposed method uses a neural network that is trained E2E to imitate the behavior of a human operator, and sequentially presents haptic guidance by capturing images using a camera attached to robot that constantly change. Therefore, the objects with considerably changing shapes such as flexible objects can be handled. Furthermore, unlike existing methods that provide deterministic haptic guidance, the proposed method predicts haptic guidance as a Gaussian distribution with uncertainty, realizing shared control that considers uncertainty in inference (i.e., providing strong guidance when AI is confident in its own inference results and weak guidance when it is not confident). The experimental results using the humanoid robot Nextage OPEN show that using the proposed system reduces the time required to complete the towel folding task by 14.5% and 17.1%, compared to the cases when the operator controls the robot alone and no uncertainty is considered, respectively.

The contributions of this letter are summarized as follows.

- To the best of our knowledge, this is the first study using imitation learning for haptic shared control.
- We proposed a method to predict the guidance as a Gaussian distribution such that the uncertainty of inference can be added to the guidance.
- The results of experiments using humanoid robot Nextage showed that the proposed system could reduce the time required for task completion by 14.5% and 17.1%, compared with the cases in which control is performed when the operator controls the robot alone and no uncertainty is considered, respectively.

II. Related Work

A. End-to-End Learning for Autonomous Robot

Several studies have explored the application of E2E learning for autonomous robot control. In the domain of automated driving, a convolutional neural network (CNN) that produces steering actions directly from the front camera image has been proposed [10]. This approach achieved excellent performance even with a small amount of training data collected for less than 100 hours of driving, demonstrating its effectiveness under various environmental conditions such as highways, regular roads, and diverse weather conditions. Furthermore, Xu et al. introduced a network that utilized large unmodified video data and previous self-driving conditions to predict a discrete sequence of vehicle self-driving behaviors, including straight ahead, stop, left turn, and right turn [11].

The success of E2E learning in autonomous driving has led to its application in complex robots with high degrees of freedom, such as humanoid robots. For example, Yang et al. combined E2E learning with an upper body humanoid robot to manipulate soft objects, which was previously considered challenging [12]. Their proposed model employed a deep convolutional autoencoder (DCAE) to extract low-dimensional image features from camera inputs, followed by a time delay neural network (TDNN) that predicted the next time-step image features and robot joint angles. During the folding movements, human operators remotely controlled the upper body robot and collected joint angle and camera image pairs as the training data. Their approach achieved a 77.8% success rate in folding task, which demonstrated the potential of E2E learning to enable complex manipulation tasks in robotics.

B. Haptic Shared Control

A fully autonomous robot refers to a machine that can perform tasks independently within its environment, relying on its sensing, planning, and acting capabilities without human intervention. Despite notable progress in automation, achieving complete autonomy in robots remains a challenge, especially when it comes to effectively handling unpredictable or unforeseen situations. For example, Kazhoyan et al. [13] demonstrated that an AI-controlled robot required 45 minutes to independently complete a table setting task, indicating the ongoing struggle of AI in performing tasks without human intervention. Consequently, most robotic applications involve human-operated or supervised robots, where a human operator provides superior situational awareness, logic, and problem-solving abilities. In the past, several robot control architectures have been developed to facilitate human interaction with partially autonomous robots. This design approach, known as "shared control," has been primarily employed in various scenarios such as remote robot operation in space or undersea exploration [14], [15], aerial robotics [16], and surgical robotics [17].

Recently, machine learning techniques have advanced significantly, and learning from demonstration (LfD) has emerged as a promising approach for training robot motions based on expert demonstrations. LfD involves learning expert behaviors during task execution and utilizing them to assist non-experts in performing similar tasks. This technique has also found applications in shared control. For instance, Luo et al. predicted an operator's actions using an autoregressive (AR) model and guided the predicted end-effector trajectory through haptic feedback, enabling smooth teleoperation [18]. Moreover, there have been attempts to model expert actions probabilistically, allowing for dynamic adjustment of system intervention based on uncertainty [19]. In their work, the end-effector trajectories performed by an expert were learned as a probability distribution, and the

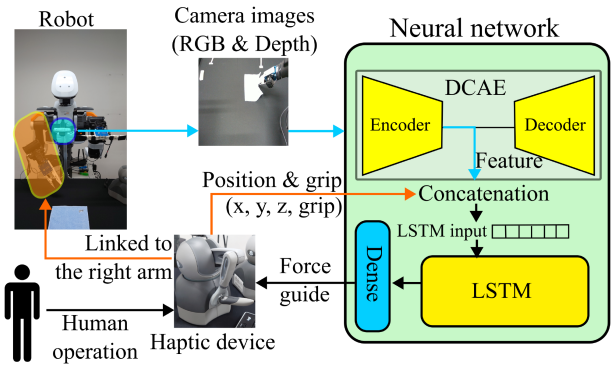

Fig. 1. Overall view of the proposed system.

intensity of haptic guidance was adjusted based on the learned trajectory's variance. Furthermore, efforts have been made to probabilistically estimate human intent [20]. This research utilized a recursive Bayesian filter to handle uncertainty in intent estimation, enabling dynamic adaptation of the behavior of the system in response to uncertainty, thereby facilitating smooth shared control. Similarly, in [21], Gaussian mixture models (GMMs) were employed to learn the trajectories performed by an expert, which were then used to generate haptic guidance, assisting in the peg-in-hole task.

While these approaches have primarily focused on generating haptic guidance using robot poses and expert control inputs, they have not fully exploited complex sensor information such as cameras. However, with the recent advancements in deep learning techniques, it has become possible to leverage such sensor information in haptic shared control. The most relevant research to our work is presented in Farraj et al. [9]. In this method, point cloud scans of all objects within the camera image are generated, creating graspable object data. Learning is then performed to determine which parts of the graspable object should be grasped, and guidance for object grasping is provided based on the learned results. When the operator is away from the object, only haptic feedback related to the danger zone is provided, while entering a predefined distance triggers feedback to guide the operator to the nearest appropriate object. This approach has achieved approximately a 31% reduction in total trajectory length and approximately a 23% reduction in grasping completion time compared to conventional control methods, specifically human-only teleoperation. While this approach incorporates sensor-based guidance generation, it does not consider uncertainty in the inference results.

III. Proposed Method

A. Overall View of the Proposed System

Fig. 1 shows an overall view of the proposed system which consists of the following four elements:

- *Haptic device:* A controller operated by the operator. The end-effector of the robot moves in coordination with the XYZ coordinates indicated by the controller. Additionally, the controller contains motors that allow it to provide haptic guidance to the operator.
- An upper-body humanoid robot.
- A camera attached to the chest of the robot and captures the surface of the table.
- *Neural Network:* Predicts the next posture of a robot from its current posture and camera image, and provides haptic guidance to the haptic device.

The haptic devices produce a four-dimensional vector $\boldsymbol{p} = (x, y, z, o)$, where the first three values correspond to the end-effector coordinates and the fourth value indicates whether the gripper attached to the robot arm is open or closed. The gripper's state, denoted as o, is typically represented by a binary value. However, for compatibility with other dimensions that employ real values, it is expressed as a real value. When the gripper is open, o takes the value 1, and when it is closed, o takes the value -1. Herein, we refer to this vector $\boldsymbol{p}$ as the position information vector. The robot arm operates in conjunction with the haptic device. Additionally, an RGB image of size 144×144 and a depth image of size 144×144 are provided from a camera installed on the chest of the robot (hereafter, we refer to the concatenated four-channel image of RGB and depth images as $\boldsymbol{I}$). The neural network consists of a DCAE, a long short-term memory (LSTM), and a fully connected layer. The DCAE performs dimensionality reduction by converting the image $\boldsymbol{I}$ into a feature vector $\boldsymbol{f}$. The position information vector at time t, $\boldsymbol{p}_t$, and the image feature vector $\boldsymbol{f}_t$ are concatenated to form the visuo-motor vector $\boldsymbol{v}_t = (\boldsymbol{p}_t, \boldsymbol{f}_t)$ which is provided to the LSTM. Then, the LSTM predicts the visuo-motor vector at time $t+1$, $\widehat{\boldsymbol{v}}_{t+1} = (\boldsymbol{p}_{t+1}, \boldsymbol{f}_{t+1})$. Finally, the haptic device provides haptic guidance according to $\widehat{\boldsymbol{v}}_{t+1}$ to assist remote operation of the robot. As described in the following subsection, the strength of the haptic guidance is dynamically adjusted based on the confidence of the LSTM to achieve smooth guidance without inhibiting human operation.

B. Confidence estimation of Neural Network Inference

In traditional time series prediction using a general LSTM, the evaluation of the reliability of the predicted value is difficult because the value is inferred deterministically. To address this issue, this study models visuo-motor vectors $\boldsymbol{v}_t$ at time t as a random variables that follows the following Gaussian distribution:

$$\widehat{\boldsymbol{v}}_t \sim \mathcal{N}\left(\widehat{\boldsymbol{v}}_t | \boldsymbol{\mu}_{\boldsymbol{v},t}, \boldsymbol{\sigma}^2_{\boldsymbol{v},\boldsymbol{t}}\right), \tag{1}$$

where $\boldsymbol{\mu}_{\boldsymbol{v},\boldsymbol{t}}$ and $\boldsymbol{\sigma}^2_{\boldsymbol{v},\boldsymbol{t}}$ are the mean and the variance of $\boldsymbol{v}_t$, respectively. Fig. 2 conceptually illustrates the time-series prediction of the proposed method. The solid black line represents the mean value of the prediction μ, and the 3σ range of the prediction is shown around the prediction value using the predicted standard deviation σ. The value of the Gaussian distribution using the standard deviation of the prediction value σ is treated as the confidence level. If the variance of the prediction value is large, the confidence level is small, and vice versa.

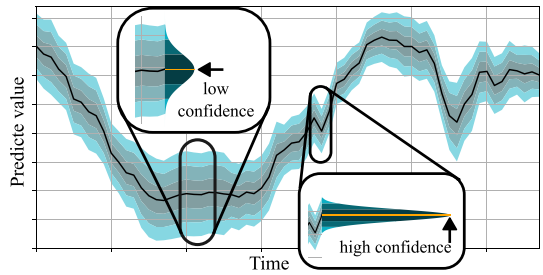

Fig. 2. Conceptual diagram of time-series prediction using the proposed method.

The following negative log likelihood is used for the loss function:

$$L\left(\boldsymbol{\mu}_{\boldsymbol{w}}, \boldsymbol{w}, \boldsymbol{\sigma}_{\boldsymbol{w}}^2\right) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\left\{\log(\sigma_{\boldsymbol{w},i}^2) + \frac{(\mu_{\boldsymbol{w},i} - w_i)^2}{\sigma_{\boldsymbol{w},i}^2}\right\}, \tag{2}$$

where, $\boldsymbol{\mu}_{\boldsymbol{w}}$ and $\boldsymbol{\sigma}_{\boldsymbol{w}}^2$ are the mean and variance of the predicted values outputted by the LSTM respectively, and, $\boldsymbol{w}$ is the ground truth signal. This study employs human demonstrations as the ground truth signal to facilitate imitation learning. The mean squared error (MSE) frequently used in machine learning corresponds to (2) when $\boldsymbol{\sigma}_{\boldsymbol{w}}^2 = \boldsymbol{E}$ (where $\boldsymbol{E}$ is a vector with N elements that are all 1). With MSE, a neural network can reduce the loss only by adjusting the parameters such that the predicted value $\boldsymbol{\mu}_{\boldsymbol{w}}$ is closer to the ground truth signal $\boldsymbol{w}$. In contrast, the loss function used in this study, (2), allows for the reduction of loss by increasing $\boldsymbol{\sigma}_{\boldsymbol{w}}^2$ when the predicted value $\boldsymbol{\mu}_{\boldsymbol{w}}$ cannot be brought closer to the ground truth signal $\boldsymbol{w}$. Therefore, by learning, the parameters reduce the loss and the predictions of the LSTM are naturally adjusted to increase the variance when the prediction is difficult. This allows us to determine the confidence in the prediction results from the variance value, and guidance force can be reduced when the variance is large and increased when the variance is small. Note that, in this letter, $\boldsymbol{\sigma}_{\boldsymbol{v}}^2$ is assumed to be a diagonal matrix.

C. Neural Network Architecture

1) Deep Convolutional Autoencoder (DCAE): We use the encoder in the DCAE to reduce the spatial resolution of the image and extract features. The DCAE is a type of autoencoder composed of an encoder and a decoder. By training it to ensure that the input to the encoder and the output from the decoder are the same, the intermediate layers can organize image features in a self-supervised manner. This enables the generation of an encoder that can compress dimensions with only the unlabeled image data.

The input image for the DCAE is a 144×144 image $\boldsymbol{I}$ that combines RGB and depth images into a four-channel image. The encoder consists of six layers: four convolutional layers each with a 2D convolutional layer, batch normalization layer, and maximum value pooling layer, and two fully connected layers with a ReLU activation layer. The output of the encoder is a 128-dimensional feature vector $\boldsymbol{f}$. Similarly, the decoder consists of six layers: two fully connected layers and four transpose convolutional layers, each with a 2D transpose convolutional layer and a batch normalization layer. The output of the decoder is a 144×144 four-channel image $\widehat{\boldsymbol{I}}$ that combines RGB and depth images. The DCAE is pre-trained to ensure that the input image to the encoder can be reconstructed by the decoder.

2) Long Short-Term Memory (LSTM): We predict the next time-step image feature vector and position information vector (i.e., visuo-motor vector) using LSTM. LSTM is a type of recurrent neural network that can perform time-series prediction using long-term data. The reasons for adopting LSTM are twofold: Firstly, the author group has prior experience in using LSTM for robot control [22]. And secondly, LSTM is well-suited for generating dynamic haptic feedback as it can generate predictions sequentially. The input to the LSTM is a 132-dimensional visuo-motor vector $\boldsymbol{v} = (\boldsymbol{p},\ \boldsymbol{f})$, which combines a four-dimensional position information vector $\boldsymbol{p}$ and a 128-dimensional image feature vector $\boldsymbol{f}$ obtained by passing a camera image $\boldsymbol{I}$ through the encoder. Given the input $\boldsymbol{v}_t$, the internal activation values $\boldsymbol{h}_t$, and cell state $\boldsymbol{c}_t$ at time-step t, the LSTM generates 500-dimensional internal activation values $\boldsymbol{h}_{t+1}$ and cell state $\boldsymbol{c}_{t+1}$ as follows:

$$(\boldsymbol{h}_{t+1}, \boldsymbol{c}_{t+1}) = \mathrm{LSTM}(\boldsymbol{v}_t, \boldsymbol{h}_t, \boldsymbol{c}_t). \tag{3}$$

The internal activation values $\boldsymbol{h}_{t+1}$ are then input to a fully connected layer:

$$\left(\boldsymbol{\mu}_{\boldsymbol{v},t+1}, \boldsymbol{\sigma}_{\boldsymbol{v},t+1}^2\right) = \mathrm{Fc}(\boldsymbol{h}_{t+1}), \tag{4}$$

which outputs two 132-dimensional vectors $\boldsymbol{\mu}_{\boldsymbol{v},t+1} = (\boldsymbol{\mu}_{\boldsymbol{p},t+1}, \boldsymbol{\mu}_{\boldsymbol{f},t+1})$ and $\boldsymbol{\sigma}_{\boldsymbol{v},t+1}^2 = (\boldsymbol{\sigma}_{\boldsymbol{p},t+1}^2,\ \boldsymbol{\sigma}_{\boldsymbol{f},t+1}^2)$ each representing the mean and variance of the next visuo-motor vectorrespectively.

D. Haptic Guidance Generation

To ensure smooth haptic guidance and minimize errors in the actions of the operator, the proposed approach utilizes a haptic device to present haptic guidance that aligns with the predicted trajectory of the robot's end-effector. The method takes into account the mean and variance values obtained from the LSTM and determines the strength of the haptic guidance by applying the following equation:

$$\boldsymbol{F}_t(\boldsymbol{x}_t, \boldsymbol{\mu}_{\boldsymbol{x},t}, \sigma_t) = \frac{1}{\sqrt{2}\sigma_t}\left(1 - \exp\left(-\frac{\sqrt{2}|\boldsymbol{x}_t - \boldsymbol{\mu}_{\boldsymbol{x},t}|}{\sigma_t}\right)\right), \tag{5}$$

where $\boldsymbol{F}_t$ is the haptic guidance presented at time t, $\boldsymbol{x}_t$ is the end-effector coordinate, $\boldsymbol{\mu}_{\boldsymbol{x},t}$ is the predicted end-effector coordinate (i.e., the predicted position vector $\boldsymbol{\mu}_{\boldsymbol{p},t}$ without gripper state), and σ_t^2 is the confidence of the prediction. The value of σ_t^2 is determined by setting a lower limit with a hyperparameter $\sigma_{\min}^2$ relative to the average prediction variance $\boldsymbol{\sigma}_{\boldsymbol{f},t}^2$ for the image feature vector at time t as follows:

$$\sigma_t^2 = \max\left(\frac{1}{N}\sum_{i=1}^{N}\sigma_{\boldsymbol{f},t,i}^2, \sigma_{\min}^2\right). \tag{6}$$

The hyperparameter $\sigma_{\min}^2$ determines the maximum value of guidance provided by the neural network and plays a crucial role

in determining the balance between human control and neural network autonomy in the overall system. If σ^2_{min} is set too small, it can lead to strong feedback that contradicts the intended human actions, thereby negatively affecting the system's operability. On the other hand, if σ^2_{min} is set too large, it results in a lack of feedback, similar to having no guidance at all, leading to longer task execution times. In our study, we explored different values for σ^2_{min} and selected those that appeared to strike a good balance between human guidance and neural network autonomy. Although we acknowledge the importance of choosing an appropriate σ^2_{min} value, our empirical findings suggest that it is not excessively sensitive and does not require fine-tuning to a great extent.

IV. Experiment

A. Experimental Setup

The experiment utilized the upper body humanoid robot Nextage OPEN from Kawada Robotics Corporation. The arms of the robot had 6-DoF; however, due to constraints on the freedom of the haptic device, we controlled only the three-dimensional spatial position with fixed end-effector rotation. The 3D Systems Touch, a pen-type haptic device capable of providing haptic feedback in 3-DoF of joint position, was used by the operator to manipulate the robot. It should be noted that if a haptic feedback device with 6 DoF, such as Virtuose 6D, were employed, it would be possible to easily incorporate haptic guidance along the rotational axis by including the rotational axis as a regression target for LSTM. An Intel RealSense Depth Camera 455 was installed near the chest of the robot for camera feedback. The EZGripper from SAKE Robotics was used as the gripper. To examine the effects of stiffness of flexible objects, a commertially available blue towel with dimensions of 20 cm $\times$ 20.5 cm and a red silk fabric handkerchief with dimensions of 23 cm $\times$ 23 cm were used. The desk was a commercially available one, and a black cloth was used to create a black background when viewed through the camera in the workspace. The use of a black cloth was to reduce reflections from fluorescent lights and similar sources. We believe that as long as the cloth used has low reflectivity and is not black, it would not affect the effectiveness of the proposed method.

B. Benchmark Task

The proposed method is evaluated through a cloth folding task in which a cloth placed on the desk is folded in half using a remote controlled robot end-effector, as well as a cloth unfolding task in which a cloth placed on the desk is unfolded that has been folded in half. In the folding task, the robot end-effector is first moved to the vicinity of the cloth, and then the cloth is grasped using the gripper. On the other hand, in the unfolding task, similar to the folding task, the robotic arm is moved close to the cloth, and the gripper is used to grasp one end of the cloth. The robot arm is then operated to unfold the cloth. The operator remotely controls the robot end-effector using a two-dimensional image from a camera displayed on a monitor. In this study, the task was performed using only the right arm of the robot. The placement of the cloch was divided into six locations in two directions horizontally and vertically, creating a total of six areas within the range of motion of the robot arm. The cloch placement is shown in Fig. 3. In the experiment, we focused on three task and cloth object combinations: folding the blue towel, folding the red handkerchief, and unfolding the red handkerchief.

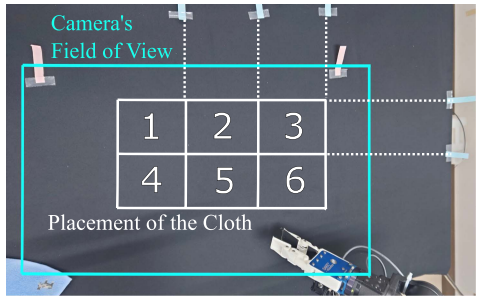

Fig. 3. Cloth placement in the experiment. The blue rectangle indicates the field of view of the robot, and the white rectangles indicate the six cloth placement locations.

C. Evaluation Metrics

In this study, three metrics were adopted for experimental evaluation: task execution time and cloth folding/unfolding accuracy, and National Aeronautics and Space Administration's task load index (NASA-TLX). For task execution time, the movement start time of the robot arm was set as 0 s, and the end time was when the robot arm returned to its initial position after folding/unfolding the cloth. The folding/unfolding accuracy is an indicator of the accuracy in achieving the desired half-fold and unfold states. The definition of accuracy is given as follows:

$$P = 1 - \frac{2S_m - S}{S} \tag{7}$$

where P represents the folding/unfolding precision, S and S_m represent the area of the unfolded and folded cloth, respectively. When the cloth is completely folded, S_m equals half of S; therefore, P equals 1. The cloth area was calculated by extracting the cloth region using OpenCV and counting the number of pixels within the region.

In order to assess workload, the NASA-TLX survey was utilized. This survey measures workload across six dimensions, namely: 1) mental demand, 2) physical demand, 3) temporal demand, 4) performance, 5) effort, and 6) frustration. Mental demand refers to the level of mental or perceptual activity needed, physical demand pertains to the degree of physical activity required, temporal demand indicates the perceived time pressure, performance captures the individual's self-evaluation of their task success, effort represents the mental and physical exertion necessary for task completion, and frustration describes the emotional state of the participant during the task, encompassing feelings of insecurity, discouragement, irritation, stress, or annoyance. Each dimension is evaluated on a scale of 0 to 100, with increments of 5 points. Lower scores are considered favorable for each dimension. In simpler terms, lower scores indicate that participants perceive lower levels of mental, physical, and

temporal demands, higher task performance, reduced effort for task execution, and decreased frustration.

D. Training Data Collection

For each of the three combinations of tasks and cloth objects mentioned earlier (folding the blue towel, folding the red handkerchief, unfolding the red handkerchief), we collected 300 samples of training data (50 samples for each of the six initial positions of the cloth). For the folding task with the blue towel, two individuals shared the data collection, with one person collecting data for positions 1 to 3 in Fig. 3, and the other was in charge of positions 4 to 6. The remaining data was collected by one person. Each trial consisted of the robot arm moving from its initial position to fold or unfold the cloth, and then returning to its initial position. Throughout the data collection process, the operators monitored the robot, rather than using a monitor. This was done to minimize the effects of dimensional information loss or feedback delay in the collection of the training data. The data was recorded at 10 frames per second.

E. Neural Network Training

We pre-trained the DCAE using the collected teacher data. We used Adam as the optimizer with a learning rate of $\alpha = 0.0005$, exponential decay rates for the first and second moments of gradients of $\beta_1 = 0.9$, $\beta_2 = 0.999$. The training was performed using mini-batch learning with 400 epochs. After extracting image features using the pre-trained DCAE to obtain the visuomotor vector, we trained the LSTM. We again used Adam as the optimizer and trained using mini-batch learning. The parameter α was set to 0.001, while the other parameters were the same as those in DCAE.

F. Experimental Results

In the experiment, three patterns were tested: (A) human-only operation, (B) haptic shared control with a constant haptic guidance strength without considering the prediction confidence of the neural network, analogous to the conventional method [9], and (C) proposed haptic shared control which dynamically adjusts the haptic guidance strength based on the prediction confidence of the neural network. The minimum variance $\sigma^2_{\min}$ in (6) was set to 0.05. For pattern (B), $\sigma_t^2 = \sigma^2_{\min}$ was set throughout the experiment.

1) Preliminary Experiment: First, to verify the effectiveness of the proposed method, we conducted an experiment with one of the individuals involved in the training data collection serving as the operator.

We performed the cloth folding/unfolding task 16 times under each condition. The initial position of the cloth was randomly selected for each of the 16 trials. Additionally, all experiments were conducted by the same person. Table I lists the task completion time and accuracy for each condition. The proposed method demonstrated the fastest task execution time without deteriorating accuracy.

TABLE I
TIME REQUIRED TO CLOTH FOLDING/UNFOLDING AND ACCURACY

Pattern	Task	Object	Task execution time (sec) mean	std	Accuracy (%) mean	std
(A)	Folding	Towel	15.6	3.9	87.1	14
	Folding	Handkerchief	12.2	1.8	97.6	3.9
	Unfolding	Handkerchief	15.8	4.6	**94.8**	9.3
(B)	Folding	Towel	14.2	3.3	84.7	8.3
	Folding	Handkerchief	12.0	2.0	98.0	2.7
	Unfolding	Handkerchief	13.5	2.8	91.0	11.8
(C): Proposed	Folding	Towel	**11.9**	**3.1**	**91.7**	**4.7**
	Folding	Handkerchief	**8.9**	**0.7**	**99.6**	**1.5**
	Unfolding	Handkerchief	**9.1**	**1.4**	92.3	**7.8**

The boldface values mean the best score in each situation considering (A), (B) and (C) case.

TABLE II
TOWEL FOLDING TIME AND ACCURACY WHEN PLACING THE TOWEL IN A POSITION NOT INCLUDED IN THE TRAINING DATA

Pattern	Task execution time (sec) mean	std	Accuracy (%) mean	std
(A)	12.5	1.3	88.2	8.4
(B)	12.3	2.0	**92.9**	**4.7**
(C): Proposed	**8.6**	**0.8**	92.0	5.4

The boldface values mean the best score in each situation considering (A), (B) and (C) case.

Additionally, an attempt was made to assess the performance by solely relying on the neural network without human intervention. Nevertheless, even though the robot managed to approach the fabric successfully, it encountered difficulties in grasping it with precision, resulting in an inability to complete the task accurately. It is worth noting that our neural network is explicitly developed for cooperation with a human operator and is not meant for complete automation. While researchers have reported successful towel folding using solely the neural network [12], the achieved success rate was 77.8%, emphasizing the importance of haptic shared control.

2) Robustness to Unseen Situation: Furthermore, we evaluated the task execution time and accuracy when the towel was placed in locations not included in the training data. Specifically, we excluded the case where the towel was placed in position 3 in Fig. 3, retrained the neural network, and performed the folding task with the towel placed in position 3 in Fig. 3. The results are shown in Table II. Similar to the previous experiment, using the proposed method (C) has successfully reduced the task execution time. This indicates that the proposed method is robust even for unknown situations.

3) Effectiveness of the Proposed Method on Inexperienced Operators: In order to determine whether the time reduction effect of the proposed method can be observed by operators with limited experience in robot operation, we conducted towel folding tasks using the robot with five individuals who were not involved in the data acquisition for training. For this experiment, we utilized the same neural network used in Section IV-F2, and positioned the towel at position 3 in Fig. 3, which was not included in the training data. The results of the experiment are presented in Table III. Since all five operators were inexperienced in robot operation, the task execution times were generally longer compared to the results shown in Table I. However, similar to the findings in Table I, it became

TABLE III
TOWEL FOLDING TIME AND ACCURACY BY INEXPERIENCED OPERATORS

pattern	Task execution time (sec) mean	std	Accuracy (%) mean	std
(A)	22.1	7.9	88.3	9.3
(B)	22.8	8.0	**90.2**	**7.9**
(C):proposed	**18.9**	**5.4**	89.1	10.0

The boldface values mean the best score in each situation considering (A), (B) and (C) case.

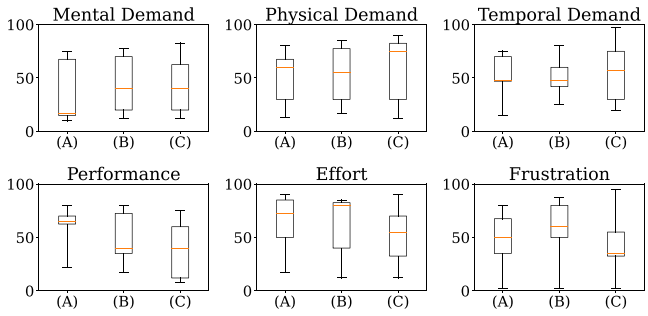

Fig. 4. Results of NASA-TLX.

evident that the proposed method (C) achieved the shortest task execution time. This confirms that the proposed method is effective even for operators who are not proficient in robot operation.

Fig. 4 displays box plots depicting the scores assigned by five operators to each item of the NASA-TLX. The introduction of the proposed method has resulted in a slight increase (deterioration) in mental, physical, and temporal demand. Conversely, performance, effort, and frustration exhibit improvement as a result of the proposed method. Notably, effort has shown a significant enhancement, indicating that the perceived level of effort required for task execution has been reduced by the proposed method.

V. DISCUSSION

As shown in Table I, the proposed haptic-shared control (C) reduced task completion time by 31.4% and 24.7% on average compared to (A) and (B), respectively. A significant difference was observed between (A) and (C) at a significance level of $p = 0.05$. Additionally, (C) exhibited smaller variance in task execution time compared to (A) or (B), indicating that the haptic guidance can reduce the variance in task execution time.

Regarding the towel folding precision, improvements of 1.5% and 3.6% were observed when compared to (A) and (B), respectively. A significant difference were found between (B) and (C) in the folding towel task, as well as between (A) and (B) in the unfolding handkerchief task at a significance level of $p = 0.05$. However, no significant difference was found for other combinations. One possible explanation of the improved accuracy in the task execution is the inclusion of task execution accuracy in the training data. The training data for towel folding had a folding accuracy of 96.6%, demonstrating precision in folding actions. Another factor is the information asymmetry between the operator and the neural network. The operator relied on visual feedback without depth information, while the neural network

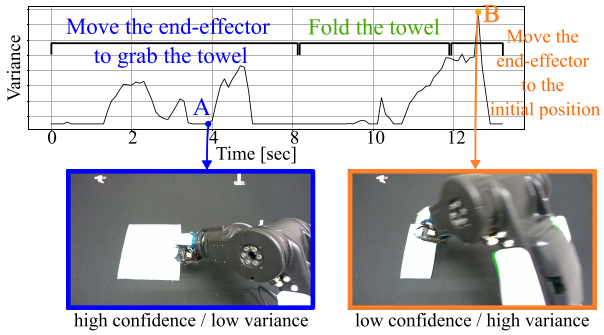

Fig. 5. Time variation of variance for a sequence and a camera image attached to the chest of the robot. At point A, shown in blue, the variance is low, the prediction is made with high confidence, and the haptic guidance is presented with high intensity. In contrast, at orange point B, the variance is high. Therefore, the prediction is made with low confidence and the haptic guidance is presented with weak intensity.

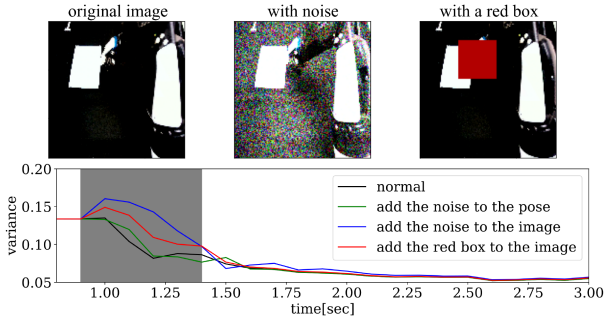

Fig. 6. Impact of interference on predicted variance. The gray shading indicates the times at which the noise and rectangle interference were introduced.

could infer depth information. The collaborative interaction between the operator and the neural network compensated for each other's limitations, enabling precise object manipulation.

Fig. 5 shows an example of the temporal evolution of the predictive variance in one sequence and the image captured by the chest-mounted camera. The predictive variance rises sharply when the towel begins to fold. This is attributed to the fact that the towel is hidden by the gripper of the robot, making prediction difficult. As mentioned before, the minimum value of variance is clipped at σ^2_{min}. Although clipping does not occur within specific time intervals, it tends to occur when the robot starts moving from the initial pose and immediately after grasping the towel, which corresponds to poses with a high occurrence frequency in the training data.

Fig. 6 presents an examination of the reliability of the prediction of the neural network. The black line represents the temporal fluctuation of predicted variance of joint angles, where higher variance indicates decreased reliability of the neural network. The green, blue depict changes in variance in predictions when uniform noise was added to the pose of the robot and image inputs, respectively. Further, the red line represents variance changes in predictions when an image with a red rectangle overlaying the towel was presented, effectively concealing the towel from the neural network. The gray shading indicates the times at which the noise and rectangle interference

were introduced. Samples of interferenced images are also provided. The results indicate that the proposed method yields high predicted variances as expected in situations where interference stimuli make inference more susceptible to failure, indicating that there is no sudden increase in guidance force. Additionally, adding noise to the image increases predicted variance, likely due to the higher dimensionality of the image feature and dominant influence on the prediction of the neural network.

Additionally, the time taken to fold towels was measured when the operator directly manipulated the robot while visually observing it, rather than using the monitor. The results revealed that when the robot was visible, the folding time was approximately half that of viewing through a screen, with the towel folding completed in about 6 seconds. This can be attributed to the limitations of depth information and the inability to freely change perspectives when viewing through the monitor. It is anticipated that future advancements, such as combining the proposed method with technologies like VR, could further enhance operability.

The results in Table III confirm the effectiveness of the proposed method for inexperienced operators in robot operation. Specifically, when compared to (B), which is analogous to the conventional methods [9], the proposed method demonstrated a notable 17.1% reduction in the time required for folding towels. T-tests showed no significant difference in task execution accuracy but a significant reduction in execution time at a significance level of $p = 0.05$, indicating that the proposed method effectively decreases task execution time without compromising accuracy.

Upon analyzing the NASA-TLX scores presented in Fig. 4, it becomes evident that there is an increase in mental workload in (B) and (C). This rise can be attributed to the requirement of collaboration with the neural network for participants in (B) and (C), unlike in (A) where they worked independently. On the other hand, performance has demonstrated improvement in both (B) and (C) compared to individual operation in (A), indicating that subjective performance is enhanced when haptic guidance is provided. It might be possible to explore haptic guidance presentations that alleviate the escalation in mental, physical, and temporal demands by tailoring the presentation method according to individual preferences.

VI. Conclusion

We proposed a haptic shared control system that dynamically adjusted the haptic guidance intensity based on the neural network confidence. Using the cloth folding/unfolding task as an example, we compared three scenarios: operator-only control, the conventional haptic shared control with a constant intensity, and the proposed method. The experimental results showed the proposed method reduced the task execution time by 14.5% and 17.1% compared to the operator-only and the existing haptic shared control system, respectively, highlightning the importance of dynamically adjusting the haptic guidance based on the uncertainty of neural network inference.

References

[1] K. Kawaharazuka, T. Ogawa, J. Tamura, and C. Nabeshima, "Dynamic manipulation of flexible objects with torque sequence using a deep neural network," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2019, pp. 2139–2145.

[2] K. Kase, K. Suzuki, P. -C. Yang, H. Mori, and T. Ogata, "Put-in-box task generated from multiple discrete tasks by aHumanoid robot using deep learning," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2018, pp. 6447–6452.

[3] N. Saito, N. B. Dai, T. Ogata, H. Mori, and S. Sugano, "Real-time liquid pouring motion generation: End-to-end sensorimotor coordination for unknown liquid dynamics trained with deep neural networks," in *Proc. IEEE Int. Conf. Robot. Biomimetics*, 2019, pp. 1077–1082.

[4] T.-C. Lin, A. U. Krishnan, and Z. Li, "Intuitive, efficient and ergonomic tele-nursing robot interfaces: Design evaluation and evolution," *ACM Trans. Hum.-Robot Interact.*, vol. 11, no. 3, pp. 1–41, Jul. 2022.

[5] L. Kaufman, *Sight and Mind: An Introduction to Visual Perception*. New York, NY, USA: Oxford, 1974.

[6] C. Ware and J. Rose, "Rotating virtual objects with real handles," *ACM Trans. Computer-Hum. Interact.*, vol. 6, no. 2, pp. 162–180, Jun. 1999.

[7] S. Zhai and P. Milgram, "Quantifying coordination in multiple DOF movement and its application to evaluating 6 DOF input devices," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 1998, pp. 320–327.

[8] D. A. Abbink and M. Mulder, "Haptic shared control: Smoothly shifting control authority?," *Cogn., Technol. Work*, vol. 14, pp. 19–28, 2011.

[9] F. Abi-Farraj, C. Pacchierotti, O. Arenz, G. Neumann, and P. R. Giordano, "A haptic shared-control architecture for guided multi-target robotic grasping," *IEEE Trans. Haptics*, vol. 13, no. 2, pp. 270–285, Apr./Jun. 2020.

[10] M. Bojarski et al., "End to end learning for self-driving cars," in *Proc. NIPS Deep Learn. Symp.*, 2016, pp. 1–9.

[11] H. Xu, Y. Gao, F. Yu, and T. Darrell, "End-to-end learning of driving models from large-scale video datasets," in *Proc. IEEE Conf. Conput. Vis. Pattern Recognit.*, 2017, pp. 2174–2182.

[12] P.-C. Yang, K. Sasaki, K. Suzuki, K. Kase, S. Sugano, and T. Ogata, "Repeatable folding task by humanoid robot worker using deep learning," *IEEE Robot. Automat. Lett.*, vol. 2, no. 2, pp. 397–403, Apr. 2017.

[13] G. Kazhoyan, S. Stelter, F. K. Kenfack, S. Koralewski, and M. Beetz, "The robot household marathon experiment," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 9382–9388.

[14] A. Douglas and H. Xu, "Real-time shared control system for space telerobotics," in *Proc. Int. Conf. Intell. Robots Syst.*, 1993, pp. 2117–2122.

[15] G. Brantner and O. Khatib, "Controlling Ocean One: Human–robot collaboration for deep-sea manipulation," *J. Field Robot.*, vol. 38, no. 1, pp. 28–51, 2021.

[16] D. Lee, A. Franchi, H. I. Son, C. Ha, H. H. Bülthoff, and P. R. Giordano, "Semiautonomous haptic teleoperation control architecture of multiple unmanned aerial vehicles," *IEEE/ASME Trans. Mechatron.*, vol. 18, no. 4, pp. 1334–1345, Aug. 2013.

[17] P.-L. Yen and T.-H. Ho, "Shared control for a handheld orthopedic surgical robot," *IEEE Robot. Automat. Lett.*, vol. 6, no. 4, pp. 8394–8400, Oct. 2021.

[18] J. Luo, D. Huang, Y. Li, and C. Yang, "Trajectory online adaption based on human motion prediction for teleoperation," *IEEE Trans. Automat. Sci. Eng.*, vol. 19, no. 4, pp. 3184–3191, Oct. 2022.

[19] F. Abi-Farraj, T. Osa, N. P. J. Peters, G. Neumann, and P. R. Giordano, "A learning-based shared control architecture for interactive task execution," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2017, pp. 329–335.

[20] S. Jain and B. Argall, "Probabilistic human intent recognition for shared autonomy in assistive robotics," *ACM Trans. Hum.-Robot Interact.*, vol. 9, no. 1, pp. 1–23, Dec. 2019.

[21] C. J. Pérez-del Pulgar, J. Smisek, V. F. Muñoz, and A. Schiele, "Using learning from demonstration to generate real-time guidance for haptic shared control," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, 2016, pp. 003205–003210.

[22] 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," *Sci. Robot.*, vol. 7, no. 65, 2022, Art. no. eaax8177.