2025 IEEE/SICE International Symposium on System Integration (SII)
January 21-24, 2025, Munich, Germany

Predicting human behavior using knowledge information in jig operation and robot collaborative action generation

Mone Tamaki[1,2], Ryoichi Nakajo[2,1], Natsuki Yamanobe[2], Yukiyasu Domae[2], and Tetsuya Ogata[1,2]

Abstract— **In human-robot collaborative tasks, learning-based models that can deal with behavior beyond the scope of human description are progressing rapidly. Deep learning is effective in capturing complex nonlinear relationships, making it valuable in scenarios with intricate interactions between the environment and tasks, such as collaborative tasks. Deep learning improves the performance by incorporating multiple information sources. Human knowledge, which is regarded as supplemental information obtained from the environment, has been shown to enhance the generalization ability of task execution when it is appropriately incorporated into the learning process for robot motion generation. Among the various models, those that utilize action labels subjectively defined by humans for robot behavior enable the robot to comprehend its own actions better, leading to higher generalization. This approach also suggests that estimating human actions contributes to predicting robot movements in human-robot collaboration (HRC). However, the performance of learning-based methods is significantly influenced by the quality of the training data. Therefore, capturing appropriate human information and integrating this information into the learning process are critical for improving the ability of the robot to learn collaborative tasks. In this study, we propose a learning model that not only provides a robot with action labels for its own behavior but also includes human action labels, encouraging the robot to respond to human actions. The optimal amount of human information to be used in learning is evaluated by adjusting the methods for defining human action labels and the quantity of human data utilized. Experiments were conducted with a task in which the robot handled the manipulation of jigs in an assembly operation involving both humans and robots. The results of the learning process suggest that estimating human behavior can assist in generating collaborative robot actions.**

I. INTRODUCTION

Robots have been increasingly adopted as a solution to worker shortages. With the increasing use of collaborative robots in industrial settings and daily environments, there is a growing need for seamless human-robot interaction. The required robotic environments have shifted from those in which robots operate independently to those in which humans and robots work collaboratively, making research in this area increasingly important [1].

Collaboration between robots and humans realizes the combination of the precision, speed, and repeatability of robots with the flexibility and cognitive skills of humans. This collaboration allows for the maintenance of system efficiency regardless of human skill or condition [1], [2]. That is, HRC complements the abilities of both humans and robots by combining their strengths to enable more efficient operations. In HRC, various control methods have been investigated to ensure that robots can collaborate with humans safely and efficiently. In model-based approaches, the actions of robots are modeled and their movements are planned and controlled based on physical laws and control theory. For example, model-based approaches such as impedance control and optimal control [3] allow robots to respond flexibly to the forces exerted by humans or the environment. This ensures human safety while improving the precision and efficiency of collaborative tasks. However, model-based methods face challenges in adapting to complex and dynamic environments.

Programming-free multimodal communication and control methods have been actively researched [4]. As one of the collaborative methods, robots adapt to changes in human actions and the environment by dynamically modifying pre-planned movements. Human and environmental recognition plays a crucial role in adapting to external changes [5]. Pre-programmed robots also have limitations in terms of their adaptability to such external changes. Because humans can exhibit behaviors beyond predefined descriptions, learning-based approaches are considered highly effective, particularly those that employ deep learning models that are capable of advanced recognition and decision-making [6]–[8]. Learning-based methods, including deep learning, are influenced by the quality of the training data, which must be processed into a suitable form for learning. The identification and integration of useful information are necessary.

Deep learning, which can model complex dynamics, has emerged as an approach for robotic manipulation tasks. Levine et al. demonstrated that the end-to-end learning of the images and motions of the robot enabled robots to perform manipulation tasks directly based on perceptual data [9]. Mukherjee et al. showed that deep learning methods are effective for robotic manipulation in HRC tasks [10].

Multitask learning with appropriate auxiliary information can improve the learning performance [11]. Human knowledge is an auxiliary modality that supports learning [12], [13]. In the generation of robotic movements, auxiliary information based on human knowledge improves the learning accuracy. For example, learning models that incorporate auxiliary information based on the subjective categorization of robot actions by humans have exhibited greater generaliz-

*This work was supported by the New Energy and Industrial Technology Development Organization (NEDO) JPNP20006

[1]M. Tamaki, R. Nakajo, and T. Ogata are with the Department of Intermedia Arts and Science, Faculty of Science and Engineering, Waseda University, Tokyo 169-8050, Japan

[2]M. Tamaki, R. Nakajo, N. Yamanobe, Y Domae, and T. Ogata are also with the National Institute of Advanced Industrial Science and Technology, Tokyo 100-8921, Japan

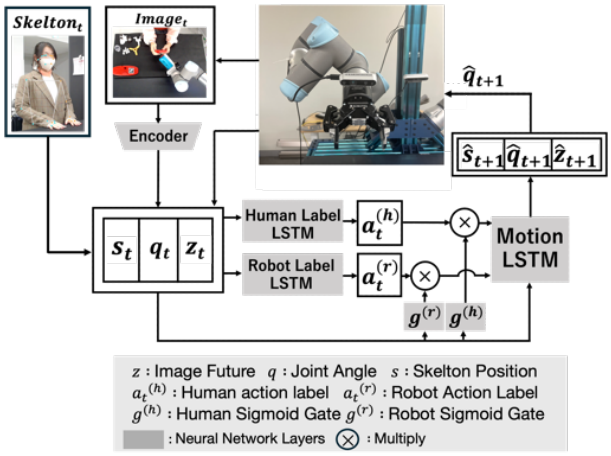

Fig. 1: Overview of our proposed model. Two label-LSTMs classify the action labels of the human and robot and a motion-LSTM predicts the future visuomotor information.

ability because their models enable robots to recognize and classify their own actions [14].

Providing auxiliary information regarding human actions allows robots to obtain information relating to all contributors to state changes, which is expected to improve the generalizability of the task execution. Furthermore, by recognizing and classifying auxiliary information related to human and robot actions, it is possible to explain the basis on which robot movements are generated. This clarifies whether the actions of the robot are spontaneous or a result of considering human conditions.

In this study, we aim to apply a motion generation model that provides robots with label information regarding human conditions, in addition to robot action labels, to HRC tasks. This approach promotes the ability of the robot to recognize and predict human actions. As shown in Fig. 1, the learning model consists of three long short-term memory (LSTM) modules: two label-LSTMs and one motion-LSTM, which estimate the actions of both humans and robots.

The two label-LSTM modules are trained to classify auxiliary information regarding both humans and robots based on current visuomotor and biometric data. By classifying the auxiliary information, the system can simultaneously account for the states of both humans and robots. The predicted action label information is then used to enhance the learning capability of the motion-LSTM.The label information classified by the label-LSTMs may vary in reliability. This is addressed using a gated mechanism that is trained to predict the confidence level of the label information. The motion-LSTM learns to predict the visuomotor and human data of the next time step from the current data and classified label information. The predicted visuomotor and human data are used to generate robot movements. Therefore, our model considers both the human and robot states in parallel, enabling the learning of collaborative tasks.

In the experiment, we used a 6-DoF robot (UR5e), along with workspace image and human data, to perform a collaborative task. The experimental task involved the collaboration of humans and a robot to assemble an airplane toy. In this study, we focused on the robot's manipulation of parts during the assembly process and collected data accordingly. The results suggested that, by labeling human actions, the robot could grasp both human and robot behaviors, enabling partial task completion in untrained positions.

II. Related Work

HRC combines the precision, speed, and repeatability of robots with the flexibility and cognitive skills of humans, enabling more efficient and adaptable operations. To achieve this, methods incorporating deep learning, which provides advanced recognition and decision-making capabilities, have increasingly been developed.

Hongyi et al. researched deep learning-based robot interfaces for HRC and integrated three methods, namely speech recognition, hand movement recognition, and body posture recognition, into a multimodal interface [6]. Mascaro et al. equipped assistive robots with the ability to recognize human-object interactions. This enabled the robots to predict the behaviors and needs of nearby humans accurately, thereby facilitating more efficient and intuitive collaboration between humans and robots [15]. Consequently, robots have been shown to generate motions that consider human conditions by estimating human actions. In this study, the proposed model generates robot actions based on the recognition and prediction of human behavior. The model used in this study demonstrates that action recognition and motion generation influence one another during learning. In terms of robot motion generation, this model allows the sequential correction of actions based on behavior recognition. Thus, by applying this model to HRC tasks, robots are expected to be able to adjust their movements continuously in response to human actions.

Deep learning has also been applied to tasks involving standalone robots. Through multitask learning with auxiliary information, the robot has access to abundant environmental data, allowing it to capture the features necessary for task execution and leading to higher precision in motion generation [16]. The modalities used for learning include visual, language, and motor information [9], [17], [18]. Human knowledge, such as feedback and human preferences, is a modality that can support the learning process [12], [13]. Stepputtis et al. conducted training by adding linguistic explanations that included the intent of the operator during demonstrations, maintaining correlations among language, perception, and motor functions. The robot considers the intent of the operator, which improves the control accuracy [19]. Tanwani et al. achieved the self-segmentation of corresponding movements by employing metric learning with video data classified by humans based on positive and negative examples [20].

Auxiliary information, which humans subjectively define to describe robot behavior, is another form of human knowledge. Kase et al. explicitly used the auxiliary information on actions derived from human knowledge to improve the

precision of robot motion generation [14]. In this model, the use of auxiliary information on actions promotes the self-recognition of the movements of the robot, thereby allowing robots to handle objects with complex and variable shapes. In this study, by applying auxiliary information to human actions, we aim to facilitate the recognition of both human and robot actions by the robot. Specifically, we propose a learning model that incorporates a training mechanism for human-related auxiliary information into a pre-existing model. By independently learning the auxiliary information for human and robot actions, the robot can distinguish between the two and recognize each. In addition, because the model can predict human behavior, the origin of the motion generation of the robot is clarified. In the experiment, we used tasks involving human intervention to train the model and explored its applicability to HRC tasks.

III. Method

This section describes a method for generating collaborative task behaviors using predictive learning, utilizing the classification of human-robot action labels.

A. Motion Generation Model Using Action Labels

The method proposed by Kase et al. [14] exhibits higher generalization in motion generation compared to conventional approaches by leveraging action labels annotated by humans for the robot motion generation. In this context, separate action label information for humans is created in addition to the action labels of the robot to facilitate the robot's understanding of human behavior.

In this study, we employ a predictive learning model that extends the GAMPL proposed by Kase et al. [14]. The learning model is illustrated in Fig. 1. GAMPL is trained to predict the motion sequence at the next time step $t+1$ from the current time step t. For application to collaborative tasks, this study introduces an additional label-LSTM to predict human action labels. Following the classification of the robot and human action labels from the current visuomotor information using two label-LSTMs, the motion-LSTM predicts the future visuomotor information.

Both the human and robot action labels $\boldsymbol{a}_t^{(h)}$, $\boldsymbol{a}_t^{(r)}$ are classified based on the image feature vectors $\boldsymbol{z}_t$, positions of the robot joints $\boldsymbol{q}_t$, and human skeletal position data $\boldsymbol{s}_t$. The output is normalized to a probability distribution using a softmax function. The gating mechanism is designed to predict the reliability of the labels and adjust the informational weight of the labels used for motion prediction, thereby enabling the effective utilization of action labels in predicting movements. The two label-LSTMs are expected to memorize the probabilities of all labels and their transitions. The gating mechanism outputs confidence scores $g^{(r)}$ and $g^{(h)}$ between 0 and 1. The levels of action label information are adjusted to $g \cdot \boldsymbol{a}$ by multiplication with the gating scores. Action labels classified from the visuomotor and human skeletal position data are fed into the motion-LSTM after passing through the gating mechanism, which are then used to predict the future visuomotor information. This process is repeated until the task is completed, allowing the model to classify action labels online without requiring label annotation during the inference time.

In HRC tasks, human actions tend to be more diverse than robot actions, and ambiguity are considered to develop in the data. The structure of the proposed model, which mitigates the negative impact of low-confidence label predictions on learning, is expected to be applicable to HRC tasks.

As shown in Fig. 1, the input information consists of the visuomotor state $\boldsymbol{y}_t$, which is a concatenation of the dimensionally reduced image feature vector $\boldsymbol{z}_t$, joint angles of the robot arm $\boldsymbol{q}_t$, and skeletal position data of the human $\boldsymbol{s}_t$ at the t-th timestep. The label-LSTM classifies the action label $\boldsymbol{a}_t$. The loss function of the label-LSTM uses the top-one label (TOL) loss proposed by Kase et al. [14].

The TOL loss is similar to the binary cross-entropy loss, but only considers the loss of the highest probability (top one) to interpret action ambiguity. The TOL loss L_{TOL} is expressed as follows:

$$L_{\mathrm{TOL}} = \frac{1}{N}\frac{1}{T}\sum_{n=0}^{N}\sum_{t=0}^{T-1}\left(-\hat{l}_{\hat{i},t}^{n}\log l_{\hat{i},t}^{n} - (1-\hat{l}_{\hat{i},t}^{n})\log(1-l_{\hat{i},t}^{n})\right) \quad (1)$$

$$\hat{i} = \arg\max_{i} l_i, \quad (2)$$

where l denotes the predicted logit outputs and $\hat{l}$ denotes the ground-truth label. i and $\hat{i}$ denote the label classes and label class with the largest logit, respectively.

The visuomotor information predicted by the motion-LSTM is optimized by minimizing the mean squared error (MSE) loss between the predicted $\hat{\boldsymbol{y}}_{t+1}^{n} = \{\hat{\boldsymbol{z}}_{t+1}^{n}, \hat{\boldsymbol{q}}_{t+1}^{n}, \hat{\boldsymbol{s}}_{t+1}^{n}\}$ and true environmental state $\boldsymbol{y}_t^n$. The MSE loss L_{MSE} is expressed as follows:

$$L_{\mathrm{MSE}} = \frac{1}{N}\frac{1}{T}\sum_{n=0}^{N}\sum_{t=0}^{T-1}\|\hat{\boldsymbol{y}}_{t+1}^{n} - \boldsymbol{y}_{t+1}^{n}\|_2, \quad (3)$$

where T denotes the total time step of the sequence, N is the index of the time-series data, t is the index set of the time steps, and n is the index set of the sequential data.

The sum of L_{TOL} and L_{MSE},

$$L = L_{\mathrm{TOL}} + L_{\mathrm{MSE}}, \quad (4)$$

is used as the training loss, and the learnable parameters are updated using the gradient descent method.

This study incorporates 2D human skeletal position data into the learning process to generate robot motions according to human movements. Human skeletal positions are also included in the label-LSTM of the robot to classify the robot actions based on human movements. The effects of human-related information for the task are evaluated by varying the amount of skeletal position data used in training.

B. Human Action Labels

Kase et al. [14] improved the generalization of robot motion generation by labeling the robot's actions as auxiliary information. Humans and robots play distinct roles in collaborative tasks owing to their different capabilities.

Human actions can be categorized into work processes and positions. In this study, the human action labels are obtained by combining two categories at different levels of abstraction. The first category is the work process, which is classified into two broad categories: "collaboration," when humans are working on a common task with the robot, and "solo," when humans are working alone. The second category involves the work positions in which the human interacts with the robot in collaborative tasks.

IV. Experiment

We designed a task in which a human and a robot collaborated to assemble an airplane toy to evaluate the proposed framework. To investigate the influence of the recognition of human behavior on the motion generation of the robot, a learning task was established in which the robot manipulated a jig for a part according to the human behavior. The goal of the task was for the robot to present parts in the correct orientation and position, as requested by the human. We employed a 6-DoF UR5e robot arm [21] equipped with a Robotiq 2F-85 gripper [22]. The task was performed using the MoveIt! motion planning framework [23]. The robot motion data generated through the motion planning were used as the training data.

As shown in Fig. 2(A), the human operator was positioned facing the robot and attached parts to the aircraft body held by the robot. The robot assisted the operator by adjusting the aircraft orientation. The specific procedure for this task was as follows:

1) The operator presented the wing part and the robot adjusted the aircraft body to align with the presented part. The operator then attached the part to the aircraft body.
2) The operator independently assembled the front wheel component.
3) The operator installed both the front and rear wheel components on the aircraft body held by the robot.

For the training data, we collected 100 samples for each of the three collaborative work positions, with the positions shifted by 80 mm, as shown in Fig. 2(B). For the evaluation, we collected 10 samples from each of the two intermediate positions between the training positions.

The joint angles of the robot, state of the gripper, RGB images, and 2D positional ions of the human skeleton were fed into the learning model. The RGB images were captured using a camera fixed above the workspace, covering the entire task area, and were resized to 96×72 for training purposes. In this task, because the operator was seated, significant movement occurred only in the upper body. The skeletal data were captured using a camera placed in front of the operator. The positions were estimated using OpenPose [24]. As shown in Fig. 2(C), the skeletal positions of the wrists, elbows, shoulders, and eyes were extracted. Three learning conditions were applied for different types of skeletal positions input to the learning model: (a) no skeletal information, (b) wrist data only, and (c) wrist, elbow, shoulder, and eye data.

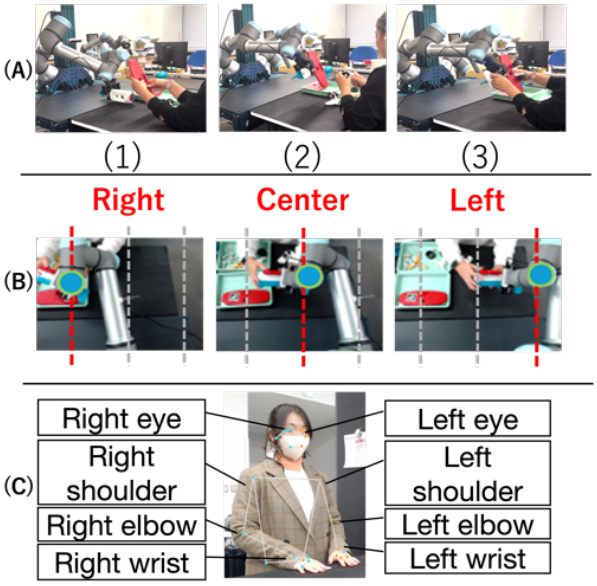

Fig. 2: Experimental settings. (A) Collaborative task. The operator switched the collaborative and independent works. (B) Acquisition position of the training data. (C) Location of human skeletal positioning information.

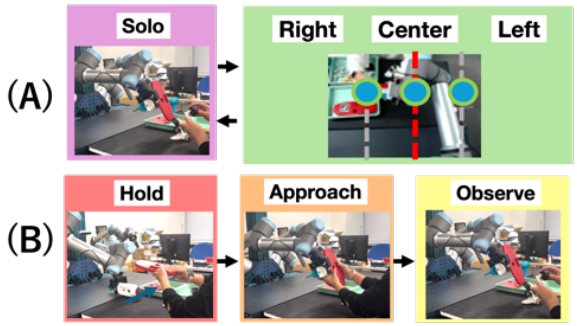

Fig. 3: Example of human and robot action label setting. (A) Human action labels were divided into independent and collaborative tasks based on the type of action. (B) Robot action labels depended on the robot motions.

Fig. 3 shows an example of the action label settings. Human actions were classified into "Solo" (independent work) and "Collaboration" (collaborative work) based on the type of task, and "Collaboration" was subdivided according to the task position. The subtasks, which were divided into solo and collaborative tasks, were repeated sequentially in the overall task. In this experiment, task steps (1) and (3) corresponded to collaborative work, while step (2) corresponded to solo work, resulting in a label transition of "Collaboration" → "Solo" → "Collaboration." As the "Collaboration" task was subdivided by position, the label transitions for a single dataset would be "Right" or "Center" or "Left" → "Solo" → "Right" or "Center" or "Left."

During training, the image data were compressed into 10 dimensions using a convolutional autoencoder. The visuomotor information $\boldsymbol{y}$ for learning consisted of concatenated

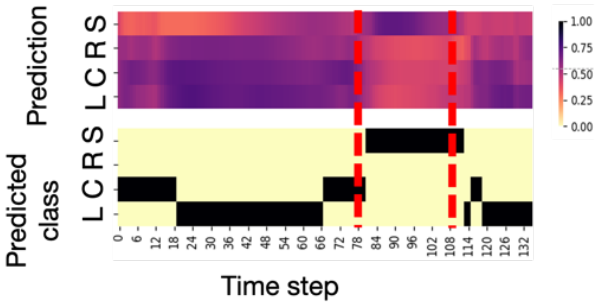

Fig. 4: Classification results for human action labels at an untrained position between the Left and Center labels.

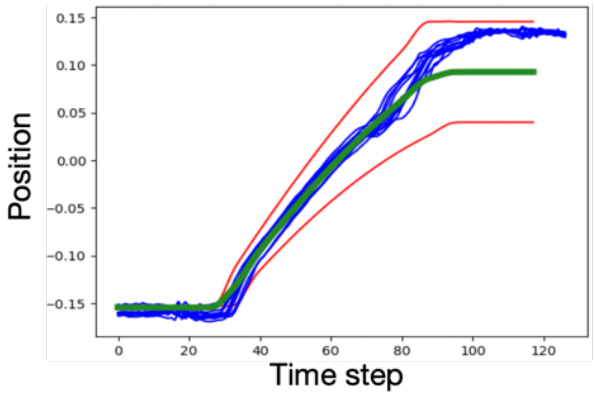

Fig. 5: Predicted tool center position of the robot at the untrained work position CL. The red lines indicate the center and left work positions. The green and blue lines indicate the expected and predicted trajectories at the CL position, respectively.

10-dimensional image features $\boldsymbol{z}$, 7-dimensional robot joint angles and gripper states $\boldsymbol{q}$, and 0-, 4-, or 16-dimensional human skeletal position information $\boldsymbol{s}$, resulting in a total of 17 to 23 dimensions. The action labels $\boldsymbol{a}^{(h)}$ and $\boldsymbol{a}^{(r)}$ were generated by the respective label-LSTMs. The dimensions of the action labels were determined based on the number of label classifications; in this experiment, the human label was 4-dimensional and the robot label was 3-dimensional.

The proposed method was evaluated based on the classification of human labels at untrained positions, accuracy of the motion predictions, and trajectories of the predicted motions. In addition, we examined the influence of human skeletal position information on learning.

V. Results and Discussions

A. Classification Results for Human Action Labels

Fig. 4 shows the results of the classified human action labels at an untrained position between the "Center" and "Left" labels. The vertical axis represents the action labels and the horizontal axis represents the time steps. The first row of the graph shows the predicted probabilities and the second row displays the label with the highest prediction probability. The dotted lines indicate the time steps for task steps (1), (2), and (3) described in Fig. 2(A). The left dotted line indicates the transition from "Solo," based on task type, to "Left," "Center," or "Right," based on task location. The right dotted line represents the reverse transition back to "Solo." In the first row of Fig. 4, "Center" and "Left" are high for task steps (1) and (3). The second row of the graph in Fig. 4 also fluctuates between these two labels. This suggests that, although the labels for untrained positions were not provided during training, the model could infer labels at untrained positions.

B. Prediction Results for Robot Motion

The results of predicting the robot motion at the untrained positions are shown in Fig. 5. This graph plots the horizontal hand position of the robot from the perspective of the operator. The vertical axis represents the position and the horizontal axis represents the time steps. The red lines indicate data for the trained positions ("Left" and "Center"), while the blue line indicates the predicted data for the untrained position between "Center" and "Left" (CL). The green line denotes the midpoint of the hand position in the training data. The hand position of the robot around the 90th time step, when the robot finished presenting the part, corresponded to the midpoint between the trained "Center" and "Left" positions. This result shows that the proposed model can present parts in the required position even if the position is unlearned. However, the predicted position tended to shift towards the trained positions over time while the human was working. The absolute error between the midpoint of the hand position in the training data and predicted result was calculated. During the 30th to 90th steps, when the robot was presenting parts, the mean error was 17.1 ± 5.47 mm. After the presentation from the operator ended and the robot was stationary from the 90th step onwards, the mean error was 48.1 ± 2.07 mm. These results indicate that challenges remain in maintaining a stable position after the presentation from the operator.

C. Skeletal Position Information

We compared the learning accuracy by adjusting the amount of skeletal information used for training under the three conditions. The accuracy of the human action label classification and motion prediction based on the amount of skeletal position information are presented in Table I. The accuracy was evaluated against the untrained work positions, between "Center" and "Left" (CL) and between "Center" and "Right" (CR). The accuracy of the human action labels was evaluated based on the rates of correct labels for the classification. The CL and CR labels were not prepared during training; thus, the action labels regarding the work positions considered the two closest positions as the correct labels. The prediction accuracy of the visuomotor information was evaluated using the MSE. We trained three different learning models for each human skeletal condition and evaluated their accuracy. Table I shows that the wrist skeletal position information could effectively contribute to

TABLE I: Skeletal information quantity and prediction accuracy

	Human label [%]		Joint	
	Mean	Variance	Mean	Variance
No skeleton	82.1	1.95×10^{-1}	18.3	22.8
Wrist only	**92.6**	6.11×10^{-3}	23.1	27.5
All skeletons	82.4	1.65×10^{-1}	29.9	62.5

the classification of human labels. However, the skeletal information did not appear to contribute to the prediction of the visuomotor information. In this experiment, both the human and robot label-LSTMs were provided with visuomotor information and skeletal position data to classify the actions while considering the human behavior. However, skeletal location information was not found to be of significant value in predicting robot movements. This result indicates that the selection of the input information for each labelled LSTM model is necessary.

VI. Conclusions and Future Work

This study investigated the applicability of a motion generation model that utilizes human-related information in collaborative tasks. The proposed model comprises two LSTMs that classify the actions of humans and robots based on human knowledge. By providing the action labels of humans based on independent and collaborative tasks, the proposed model can classify the work position and predict collaborative motions in assembly tasks. In collaborative tasks, the generation of adaptive motions by robots is essential. In this experiment, the label transitions were fixed for simplification. However, these processes often involve multiple potential pathways. Murata et al. demonstrated the adaptability to situational changes in human-machine collaborative tasks by dynamically adjusting the task goal states based on the principle of minimizing prediction errors [25]. Our proposed model may be adaptable to situational changes by incorporating dynamic adjustments to the goal states. Future work will focus on developing a more intuitive method for creating datasets using a remote control to build a learning method that integrates the experience of adaptive tasks.

ACKNOWLEDGMENT

This work was conducted at the Artificial Intelligence Research Center of the National Institute of Advanced Industrial Science and Technology, and is based on the results obtained from the project JPNP20006 commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References

[1] V. Villani, F. Pini, F. Leali, and C. Secchi, "Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications," *Mechatronics*, 2018.

[2] E. Matheson, R. Minto, E. G. G. Zampieri, M. Faccio, and G. Rosati, "Human-robot collaboration in manufacturing applications: A review," *Robotics*, vol. 8, p. 100, 2019.

[3] E. Todorov and M. I. Jordan, "Optimal feedback control as a theory of motor coordination," *Nature neuroscience*, vol. 5, no. 11, pp. 1226–1235, 2002.

[4] L. Wang, S. Liu, H. Liu, and X. Wang, "Overview of human-robot collaboration in manufacturing," pp. 15–58, 2020.

[5] Q. Xiong, J. Zhang, P. Wang, D. Liu, and R. X. Gao, "Transferable two-stream convolutional neural network for human action recognition," *Journal of Manufacturing Systems*, vol. 56, pp. 605–614, 2020.

[6] H. Liu, T. Fang, T. Zhou, Y. Wang, and L. Wang, "Deep learning-based multimodal control interface for human-robot collaboration," *Procedia Cirp*, vol. 72, pp. 3–8, 2018.

[7] S. H. Choi, K.-B. Park, D. H. Roh, J. Y. Lee, M. Mohammed, Y. Ghasemi, and H. Jeong, "An integrated mixed reality system for safety-aware human-robot collaboration using deep learning and digital twin generation," *Robotics and Computer-Integrated Manufacturing*, vol. 73, p. 102258, 2022.

[8] Z. Liu, Q. Liu, W. Xu, Z. Liu, Z. Zhou, and J. Chen, "Deep learning-based human motion prediction considering context awareness for human-robot collaboration in manufacturing," *procedia cirp*, vol. 83, pp. 272–278, 2019.

[9] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *Journal of Machine Learning Research*, vol. 17, no. 39, pp. 1–40, 2016.

[10] D. Mukherjee, K. Gupta, L. H. Chang, and H. Najjaran, "A survey of robot learning strategies for human-robot collaboration in industrial settings," *Robotics and Computer-Integrated Manufacturing*, vol. 73, p. 102231, 2022.

[11] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[12] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1238–1274, 2013.

[13] A. Najar and M. Chetouani, "Reinforcement learning with human advice: a survey," *Frontiers in Robotics and AI*, vol. 8, p. 584075, 2021.

[14] K. Kase, C. Utsumi, Y. Domae, and T. Ogata, "Use of action label in deep predictive learning for robot manipulation," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 13 459–13 465.

[15] E. V. Mascaro, D. Sliwowski, and D. Lee, "HOI4ABOT: Human-object interaction anticipation for human intention reading assistive roBOTs," in *7th Annual Conference on Robot Learning*, 2023. [Online]. Available: https://openreview.net/forum?id=rYZBdBytxBx

[16] M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," in *2019 International conference on robotics and automation (ICRA)*. IEEE, 2019, pp. 8943–8950.

[17] 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.

[18] P. Florence, L. Manuelli, and R. Tedrake, "Self-supervised correspondence in visuomotor policy learning," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 492–499, 2019.

[19] S. Stepputtis, J. Campbell, M. Phielipp, S. Lee, C. Baral, and H. Ben Amor, "Language-conditioned imitation learning for robot manipulation tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 13 139–13 150, 2020.

[20] A. K. Tanwani, P. Sermanet, A. Yan, R. Anand, M. Phielipp, and K. Goldberg, "Motion2vec: Semi-supervised representation learning from surgical videos," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 2174–2181.

[21] Universal Robots. (2024) UR5e Lightweight, versatile cobot. [Online]. Available: https://www.universal-robots.com/products/ur5-robot/

[22] Robotiq. (2024) 2F-85 and 2F-140 Grippers. [Online]. Available: https://robotiq.com/products/2f85-140-adaptive-robot-gripper

[23] S. Chitta, I. Sucan, and S. Cousins, "Moveit![ros topics]," *IEEE Robotics & Automation Magazine*, vol. 19, no. 1, pp. 18–19, 2012.

[24] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, 2021.

[25] S. Murata, W. Masuda, J. Chen, H. Arie, T. Ogata, and S. Sugano, "Achieving human–robot collaboration with dynamic goal inference by gradient descent," in *Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part II 26*. Springer, 2019, pp. 579–590.