Motion Prediction for Human-Robot Collaborative Tasks Using LSTM

Kaihei Okada [1 *], Tokuo Tsuji [1 **], Tatsuhiro Hiramitsu [1], Hiroaki Seki [1], Toshihiro Nishimura [1], Yosuke Suzuki [1] and Tetsuyou Watanabe [1]

[1] *Kanazawa University, Kakuma-machi, Kanazawa 920-1192, Japan*

*Email: *kaihei2112@stu.kanazawa-u.ac.jp, **tokuo-tsuji@se.kanazawa-u.ac.jp*

Abstract

This study proposes an assistive robot system to reduce caregiving burdens in an aging society by supporting impaired body movements. The system focuses on bimanual tasks, such as pouring a drink from a bottle into a cup. Using 3D skeletal data excluding the impaired left hand, a deep learning model (LSTM) predicts the motion stages and 3D positions of the left hand, and the robot performs the substitute motions. The system uses data from multiple users to show its potential for improving patient independence and reducing caregiver workload.

Keywords: Human motion prediction, LSTM, Three-dimensional human skeleton, Caregiving robots

1. Introduction

1.1. Background and Objective

Japan is facing a serious issue of labor shortages in caregiving and nursing due to its rapidly aging society. To address this problem, the development of assistive robots has garnered significant attention. The objective of this study is to support daily life activities for patients whose body parts are paralyzed or missing due to aging, accidents, or illnesses by utilizing assistive robots. Specifically, the focus is on tasks that require the use of both hands. For instance, in the case of a patient whose left hand is paralyzed, this study proposes a mechanism to predict the movements of the left hand using a deep learning model based on the movements of other body parts and to replicate these movements with a robot. This approach is generalizable, aiming to construct a system that can predict the movement of specific body parts (e.g., right hand, left hand, right foot, left foot) from the movements of other parts and supplement these movements with a robot. This system is expected to expand the range of possible activities for patients and reduce the burden on caregivers.

1.2. Research Targets and Challenges

This study focuses on an everyday task that requires the use of both hands: "pouring a drink from a bottle into a cup and drinking it." This task includes a sequence of motions such as opening the bottle cap, pouring the drink into the cup, drinking from the cup, and closing the cap. The movements of the left hand during these motions are classified into seven stages: stopping, approaching, grasping, holding, pouring, releasing, and leaving. However, since the motions of stopping and holding are treated as the same stage, the movements of the left hand are ultimately classified into six stages. These stages are linked to the movements of the right hand, and their cooperation is an important feature. Furthermore, as shown in Fig. 1, the task and the movement stages of the left hand are visually explained. The motions of grasping and releasing are momentary and therefore difficult to predict. To address this challenge, these motions are reformulated as a binary classification problem regarding whether the hand is open or closed. Consequently, the movement stages are treated as four multi-class classification problems (stopping, approaching, pouring, leaving), while the hand state is treated as a binary classification problem (open, close).

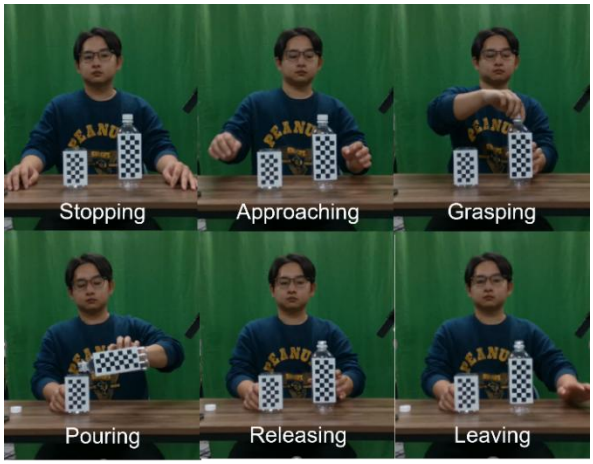

Fig. 1 Example of motion stages of left hand

1.3. Research Approach

The goal of this study is to classify the movement stages and hand states of the left hand and predict its three-dimensional movement. To achieve this, data obtained from the movements of the right hand and other body parts are utilized, and a deep learning model capable of time-series prediction is employed to estimate the movements of the left hand. Based on these estimations, the robot generates movements that replicate those of the

left hand. The proposed system is applicable to many tasks required in daily life and aims to support the lives of the elderly and those with physical limitations. Additionally, the system is expected to reduce the workload of caregivers and contribute to the efficiency of caregiving settings.

2. Related Work

As the background of this research, we review technologies related to prosthetics and assistive devices, human keypoint detection technologies, and time-series prediction techniques. By organizing these topics, we aim to clarify the positioning and novelty of this study.

2.1. Prosthetics for Replacing Missing Limbs

Prosthetics come in various types, including cosmetic prostheses, work-oriented prostheses, active prostheses, and myoelectric prostheses. Among these, active prostheses and myoelectric prostheses are primarily used to assist with daily tasks.

Active prostheses use a harness attached to the shoulder or body to control movement, enabling motions such as elbow flexion and hand opening and closing. This technology allows for simple movements as well as motions requiring both hands. In contrast, myoelectric prostheses detect the weak electrical currents (myoelectric signals) generated by muscle contractions to control the opening and closing of the hand. The primary advantages of myoelectric prostheses include high gripping strength, the ability to open and close the hand without changing the prosthesis' position, and reduced discomfort due to the absence of a harness.

In recent years, advanced myoelectric prostheses have been developed. For example, the i-Limb Ultra [3], developed by Össur, features individually actuated fingers, and supports various grip patterns. However, myoelectric prostheses also face challenges such as high costs and the difficulty of learning how to operate them. This study draws inspiration from such technologies to propose a novel support method for addressing physical impairments such as limb loss or paralysis.

2.2. Assistive Devices for Supporting Paralyzed Limbs

Assistive devices for patients with paralyzed limbs have been extensively studied. For instance, assistive hands are devices designed to support the movement of paralyzed hands or arms by leveraging the patient's residual muscle strength. These devices are also used in rehabilitation programs aimed at restoring muscle strength [4].

This study aims to develop a system that can support both patients with limb loss and those with paralysis. This feature stems from the proposed hardware's independence from the extent of physical damage. In this respect, our approach differs from conventional assistive devices and prosthetics.

2.3. Keypoint Detection for Human Motion Analysis

Keypoint detection has garnered attention as a technology for analyzing human motion. OpenPose [1] is widely used for real-time detection of 2D keypoints. By detecting keypoints for the entire body, including hands, face, and torso, OpenPose has been applied in motion analysis and rehabilitation support systems.

This study utilizes OpenPose to detect the patient's body movements and predict the motion of missing or paralyzed limbs based on this data. By integrating keypoint detection technology into a novel life-support system, we aim to enhance assistive capabilities.

2.4. Time-Series Prediction with Deep Learning

In the field of motion prediction, deep learning methods for handling time-series data have been extensively researched. Notably, Long Short-Term Memory (LSTM) networks excel in learning long-term dependencies and have demonstrated high performance in time-series prediction tasks [2]. LSTM has been applied in various domains, including motion data analysis and action recognition.

This study employs LSTM to predict the motion of missing or paralyzed limbs based on movement data from healthy body parts. By doing so, we aim to accurately reproduce the patient's physical movements and enable seamless assistance by life-support robots.

3. Proposed Alternative System for Human Motion

3.1. Overview of the Proposed System

The proposed system consists of three cameras and a robot, arranged as illustrated in Fig. 2. A participant performs a task while being recorded by the cameras. Each camera detects 2D skeletal keypoints of the human. Using stereo disparity, the 3D skeletal structure is reconstructed. A deep learning model, specifically an LSTM, is employed to predict the 3D positions of the left

hand and left elbow, as well as the motion stages, based on the 3D positions of other skeletal keypoints. The robot's motions are then generated based on the predictions. Fig. 3 provides an overview of the prediction method.

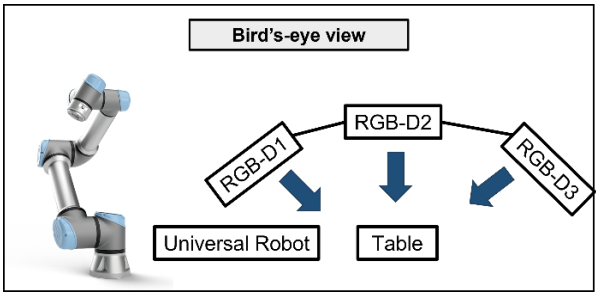

Fig. 2 Overview of the system

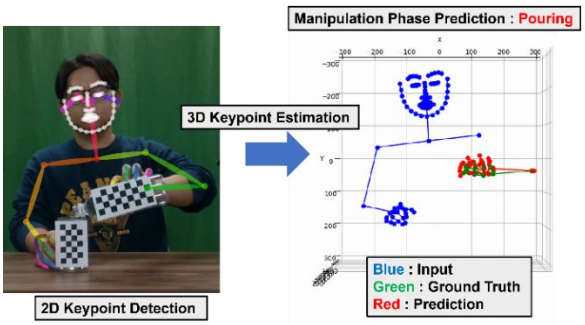

Fig. 3 Overview of the method

3.2. 2D Skeletal Detection

For videos captured by the cameras, OpenPose is used to detect human keypoints. The keypoints include 25 points for the body, 21 points for each hand, and 70 points for the face, totaling 137 keypoints. OpenPose is used because eye and hand movements are important for motion prediction. Additionally, the positions and orientations of objects are critical for the task. Checkerboard patterns are attached to objects for detection. Fig. 4 shows the detected keypoints of the human and objects during the task.

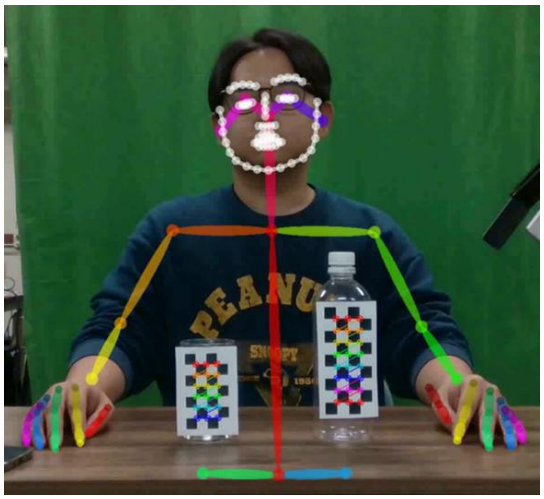

Fig. 4 Detection of keypoints and objects

3.3. 3D Skeletal Estimation

3.3.1. Stereo Calibration and Homogeneous Transformation Matrix

Camera calibration is performed to determine the projection matrices P_1 and P_2 for each camera. These matrices represent the relationship between points in 3D space and their corresponding positions in the camera image planes. In addition, the homogeneous transformation matrix H_{12} is obtained by stereo calibration. The homogeneous transformation matrix represents the relationship between cameras.

3.3.2. 3D Keypoint Estimation using DLT Method

Assume that a 3D point $X = (x, y, z, 1)^T$ is observed on the image planes of two cameras. The corresponding 2D projections are $U_1 = (u_1, v_1, 1)^T$ and $U_2 = (u_2, v_2, 1)^T$. Using the projection matrices P_1 and P_2, the relationship between these 2D points and the 3D point can be expressed as:

$$U_1 = \alpha P_1 X$$
$$U_2 = \beta P_2 X$$

where α *and* β are scale factors. Since the 2D points and their corresponding 3D projections are collinear, the cross product between U_1 and $P_1 X$ must be zero, which leads to a set of linear equations. Similarly, a similar set of equations is derived for the second camera. By combining these equations from both cameras, we obtain a system of linear equations. To solve for the 3D point X, Singular Value Decomposition (SVD) is used. The SVD decomposition of the matrix provides the optimal solution that minimizes the error due to observation noise. The smallest singular value corresponds to the column vector in V that represents the 3D point X. By applying this method to all keypoints across all frames, time-series data is obtained.

3.4. Time Series Prediction Using LSTM

In this study, we employ the deep learning model Long Short-Term Memory (LSTM) to accomplish the following tasks. First, we predict the left hand and left elbow keypoints and the position and posture of an object from the keypoints of the left hand and left elbow excluding those keypoints. Then, we predict the motion stage and the state label of the left hand. The following symbols are defined for this study. L_t represents the left hand and left elbow keypoints at frame t, R_t represents the keypoints of the left hand and left elbow excluding these at frame t, O_t represents the position and posture of an object, P_t represents the motion stage label of the left hand, and H_t represents the state label of the left hand.

Using information from up to k frames before the current frame, the left hand and left elbow keypoints, L_t can be predicted using the following model:

Tabel 1 Metrics for validation data

Metrix	RMSE[mm]	Acc(Motion)[%]	F1(Motion)	Acc(Hand)[%]	F1(Hand)
Keypoint Model	52.3	N/A	N/A	N/A	N/A
Label Model	N/A	0.759	0.697	0.820	0.817
Keypoint Label Model	**40.5**	**0.906**	**0.819**	**0.959**	**0.958**

$$L_t = LSTM(R_{t-k+1}, \ldots, R_t)$$

Furthermore, if the initial position and posture of the object, O_0, is known, we predict the values for the next frame, as follows:

$$L_t, O_{t+1} = LSTM(R_{t-k+1}, \ldots, R_t, O_{t-k+1}, \ldots, O_t)$$

Additionally, we consider a model that predicts P_t and H_t. This model can be expressed as follows:

$$L_t, O_{t+1}, P_t, H_t = LSTM(R_{t-k+1}, \ldots, R_t, O_{t-k+1}, \ldots, O_t)$$

Fig. 5 shows the structure of the LSTM model.

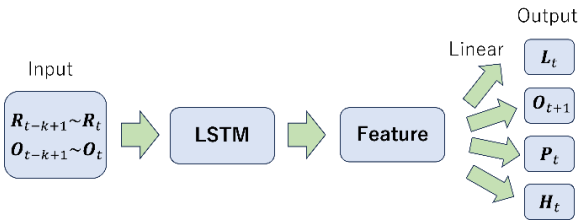

Fig. 5 Structure of our model

Furthermore, models that predict only L_t and O_{t+1} are referred to as the "**Keypoint Model**," those that predict only P_t and H_t are the "**Label Model**," and models that predict all outputs are referred to as the "**Keypoint Label Model**." Results from these models will be compared.

4. Experiment and evaluation

4.1. Experimental Setup and Dataset Creation

In this study, three RGB-D cameras (RealSense D435) were utilized, focusing exclusively on color images. Additionally, a robot arm (Universal Robot e5) was employed to assist with tasks. During data collection, the participants were seated on a chair and instructed to perform a series of tasks on a table: opening a bottle cap, pouring water into a cup, drinking water from the cup, and closing the cap. Each sequence of tasks generated one set of time-series data. A total of 70 datasets were recorded from five participants.

The collected data was split into 90% for training and 10% for validation. The time-series model was trained using the training data, and its performance was evaluated using the validation data. This evaluation aimed to quantitatively demonstrate the effectiveness of the proposed system.

4.2. Model Training

For model training, the loss function for 3D keypoint position estimation was defined as Mean Squared Error (MSE) Loss, while Multi-Class Cross Entropy Loss was used for task stage classification. The total loss function was weighted as follows:

$$Loss = 100 \times MSE\ Loss + Cross\ Entropy\ Loss$$

The Adam optimizer was employed for training, with the number of epochs set to 100. This configuration enabled efficient and effective model training.

4.3. Evaluation Metrics for Validation Data

Table 1 summarizes the average results for all validation data. The comparison includes the model predicting only the 3D positions of keypoints, the model predicting only motion stage and hand state labels, and the model predicting both (referred to as the **Keypoint Label Model**). The 3D positions of keypoints are evaluated using RMSE, while motion stage and hand state labels are evaluated using accuracy and F1 Score.

As shown in Table 1, the **Keypoint Label Model** achieved the best performance across all metrics. This result suggests that predicting labels enhances the accuracy of keypoint predictions and improving keypoint predictions also benefits label prediction accuracy. This synergy is attributed to the ability of the LSTM to effectively learn features from sequential data that are relevant for predicting both keypoints and labels. Since keypoints and labels are interrelated, they provide valuable information for each other's predictions. Moreover, the multitask learning approach enhanced the generalization capability of the model, mitigating the risk of overfitting. These findings align with previous

research on multitask learning, demonstrating its effectiveness in improving overall model performance.

4.4. Confusion Matrices for Validation Data

Fig. 6 presents the confusion matrices for hand state and motion stage label predictions, respectively, generated by the **Keypoint Label Model**. The hand state labels are predicted with high accuracy. Similarly, the motion stage labels are also predicted with high accuracy overall, although some labels exhibit lower performance. Specifically, the model tends to misclassify the "approaching" stage as the "stopping" stage and the "leaving" stage as the "stopping" stage.

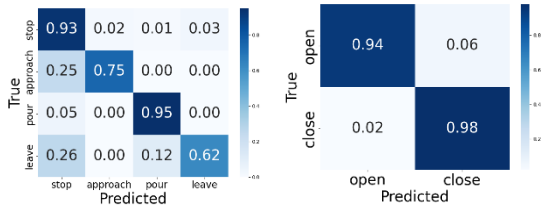

Fig. 6 The confusion matrix of Motion stages and hand state

To investigate these misclassifications, the temporal flow of labels was visualized. Fig. 7 shows the ground truth labels and predictions of motion stages over time for Validation Data 2. Around frame 350, it can be observed that the transition to the "approaching" stage occurs prematurely, leading to the misclassification of the "approaching" stage as the "stopping" stage. Similarly, around frame 200, the preceding "stopping" stage causes misclassifications in the subsequent "leaving" stage. These observations suggest that improving the timing of stage transitions could enhance prediction accuracy.

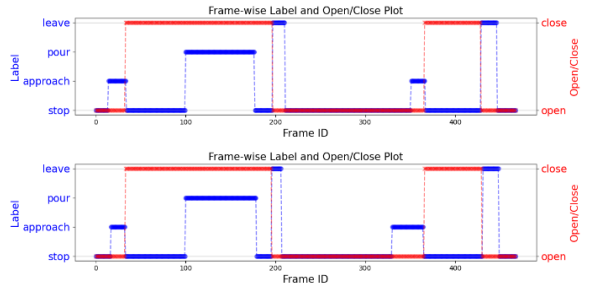

Fig. 7 The top figure shows the ground truth, while the bottom figure represents the predictions, each plotting the flow of the motion stages.

4.5. 3D Keypoint Prediction for Validation Data

Fig. 8 visualizes sampled predictions of 3D keypoints over time made by the **Keypoint Label Model**. The blue points represent the input, green points represent the ground truth, and red points represent the predictions. Although the actual predictions involve 3D positions, they are projected to 2D for visualization purposes.

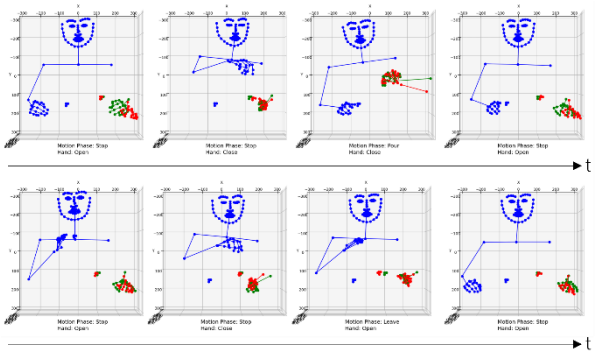

Fig. 8 The predicted 3D human skeleton

As shown in Fig. 8, the predicted keypoints closely match the ground truth, demonstrating the high accuracy achieved by the **Keypoint Label Model** in 3D keypoint prediction. However, certain keypoints, such as the initial position of the left hand or the elbow position during pouring motions, are inherently difficult to predict accurately, leading to lower performance for these specific cases.

These results highlight the strengths of the **Keypoint Label Model** in simultaneously predicting keypoints and labels with high accuracy, while also identifying potential areas for further improvement in challenging scenarios.

5. Conclusion

In this study, we utilized an LSTM-based model to predict the 3D positions of the left hand and elbow, as well as the motion stages, using data from regions other than the left hand and elbow. For the experiments, a dataset of 70 samples was created using five participants. Evaluation results on the validation data demonstrated that the model predicting both the 3D positions and motion stages simultaneously achieved higher accuracy compared to models predicting them separately. In future work, we aim to investigate the model's capability to handle an increased variety of motion types. Additionally, we plan to evaluate the generalization performance of the model on data from individuals outside the current dataset.

References

1. Hochreiter, S. (1997). Long Short-term Memory. Neural Computation MIT-Press.
2. Cao, Z., Simon, T., Wei, S. E., & Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7291-7299).
3. Ossur, "i-Limb Ultra." Accessed: Dec. 26, 2024. [Online]. Available: https://www.ossur.com/en-gb/prosthetics/arms/i-limb-ultra

4. H. Ishisaki, K. Tabata, T. Tsuji and T. Watanabe, "Hand Grasping Assist Glove Combining Exoskeleton Structure and Pneumatically Driven Actuator," 2023 IEEE/SICE International Symposium on System Integration (SII), Atlanta, GA, USA, 2023, pp. 1-5, doi: 10.1109/SII55687.2023.10039272.

Authors Introduction

Mr. Kaihei Okada

He received his B.S. degree in engineering from Kanazawa University, Japan, in 2023. He is currently a master's degree student in the Division of Frontier Engineering, Kanazawa University. His research interest includes robot vision and life support robots.

Dr. Tokuo Tsuji

He received his BS, MS, and doctoral degrees from Kyushu University in 2000, 2002, and 2005, respectively. He worked as a research fellow of Graduate School of Engineering, Hiroshima University, from 2005 to 2008. He worked as a research fellow of Intelligent Systems Research Institute of National Institute of Advanced Industrial Science and Technology (AIST) from 2008 to 2011. From 2011 to 2016, he worked as a research associate at Kyushu University. From 2016, he has been working as an associate professor at Institute of Science and Engineering, Kanazawa University. His research interest includes multifingered hand, machine vision, and software platform of robotic systems.

Dr. Tatsuhiro Hiramitsu

He is assistant professor of Institute of Science and Engineering, Kanazawa University. He received Dr E. degrees from school of engineering, Tokyo Institute of Technology, Japan, in 2019. His research interest is in the soft structure mechanisms for robotic systems. He is a member of the Japan Society of Mechanical Engineers (JSME), the Robotics Society of Japan (RSJ), and Institute of Electrical and Electronics Engineers (IEEE).

Dr. Hiroaki Seki

He received his Ph.D. in precision machinery engineering from the University of Tokyo in 1996. He is currently a professor of Institute of Science and Technology in Kanazawa University. His research interests include novel mechanism and sensor system in robotics and mechatronics.

Dr. Toshihiro Nishimura

He received the B.S., M.S., and Ph.D. degrees in mechanical engineering from Kanazawa University, Kanazawa, Japan, in 2016, 2018, and 2019, respectively. He was a Researcher of industrial robots with FANUC Corporation, from 2018 to 2021. He is currently an Assistant Professor with the Faculty of Frontier Engineering, Institute of Science and Engineering, Kanazawa University. His research interests include robotic hands, mechanism design, soft robotics, and 3D printer.

Dr. Yosuke Suzuki

He received the B.Eng., M.Eng., and Ph.D. degrees in engineering from the Tokyo Institute of Technology, Tokyo, Japan, in 2005, 2007, and 2010, respectively. He is currently an Associate Professor with the Faculty of Mechanical Engineering, Institute of Science and Engineering, Kanazawa University, Kanazawa, Japan. His research interests include tactile and proximity sensors, robotic grasping, and distributed autonomous systems.

Dr. Tetsuyou Watanabe

He is a professor with Kanazawa University. He received the B.S., M.S., and Dr.Eng. degrees in mechanical engineering from Kyoto University, Kyoto, Japan, in 1997, 1999, and 2003, respectively. From 2003 to 2007, he was a Research Associate with the Department of mechanical Engineering, Yamaguchi University, Japan. From 2007 to 2011, he was an assistant professor with Division of Human and Mechanical Science and Engineering, Kanazawa University. From 2011 to 2018, he was an associate professor with Faculty of Mechanical Engineering, Institute of Science and Engineering, Kanazawa University. Since 2018, he has been a professor with Kanazawa University. From 2008 to 2009, he was a visiting researcher at Munich University of Technology. His current research interests include robotic hand, grasping, object manipulation, medical and welfare sensors, surgical robots, and user interface. He got several awards including best paper award at Transactions of the Society of Instrument and Control Engineers and World Robot Summit Second Prize of World Robot Challenge Industrial Robotics Category Second Prize of World Robot Challenge Industrial Robotics Category.