End-to-End Learning-Based Non-Verbal Behavior Generation of Social Robots

Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, and Jaehong Kim
Artificial Intelligence Research Laboratory
Electronics and Telecommunications Research Institute (ETRI)
Daejeon, Republic of Korea
{wrko, minsu, leejy, jhkim504}@etri.re.kr

***Abstract*—In order for users to feel familiar with social robots, it is important for social robots to generate non-verbal robot behaviors, such as *handshakes*. However, the traditional approaches of reproducing pre-coded motions allow users to easily predict the robot's reaction, giving the impression that the robot is a machine and not a real agent. To enable social robots to learn multiple human-like behaviors from human-human interactions, we proposed an end-to-end learning-based behavior generation method. The Seq2Seq architecture consisting of two long short-term memory units was adopted. One is for encoding user behavior and the other is for generating the next robot behavior. To demonstrate the effectiveness of our method, two experiments were performed using a humanoid robot, Pepper, in a simulated environment. Experimental results showed that the robot can generate five social behaviors, i.e. *bow*, *stand*, *handshake*, *hug*, and *block face* corresponding to user behavior, and adjust its behavior according to the user's posture.**

***Index Terms*—social robot, social behavior, behavior generation, end-to-end learning, deep learning**

I. Introduction

To provide effective and enjoyable human-robot interaction, it is important for social robots to perform human-like non-verbal behaviors [1]. Social robots should *greet* users when they come home, *shake hands* when they reach out a hand, and *hug* them when they cry. To do this, the robots needs be able to recognize the user's posture, detect their behaviors, and generate natural motions within an appropriate time. Although the frameworks for learning non-verbal behaviors from human-human interaction have been proposed [2], they usually focus only on one behavior, such as *handshake*.

In this paper, we proposed a neural network structure that can learn multiple non-verbal behaviors in an end-to-end way. The model consists of two long short-term memory (LSTM) units [3], one for encoding the previous user poses, and the other for generating the next robot behavior. The weights of the model were trained using a human-human interaction dataset, *AIR-Act2Act* [4]. The experiments were carried out using a humanoid robot, Pepper, in a simulated environment.

This paper is structured as follows. In addition to this introductory section, Section II describes the proposed method for generating non-verbal behaviors of social robots. Section III validates our method using a human-human interaction dataset. Section IV presents conclusions.

II. Proposed Method

Figure 1 shows the overview of our proposed method for behavior generation using Seq2Seq architecture [5]. It consists of two LSTM units [3]. One is an LSTM encoder that encodes user behavior $\boldsymbol{U}$ into a vector $\boldsymbol{z}$, and the other is a LSTM decoder that generates the next robot behavior $\bar{\boldsymbol{R}}$ corresponding to the current robot pose $\boldsymbol{R}$ and the $\boldsymbol{z}$. User behavior $\boldsymbol{U}$ is a series of user poses, each of which is represented as $\boldsymbol{P}_t = [\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_k]$, where $\boldsymbol{p}_k = (x_k, y_k, z_k)$ indicates the 3D coordinates of the k-th body joint, such as shoulder, elbow. $\boldsymbol{z}$ is the extracted feature vector of user behavior, values of which are sampled from the probability distributions with means $\boldsymbol{\mu}$ and variances $\boldsymbol{\sigma}$. Robot behavior $\bar{\boldsymbol{R}}$ is a series of robot poses, each of which is represented as robot joint angles, such as pitches and rolls of shoulders, yaws and rolls of elbows. Note that, in the proposed model, teacher forcing and generative adversarial network techniques were used, but we omit the detailed explanations due to space limitation.

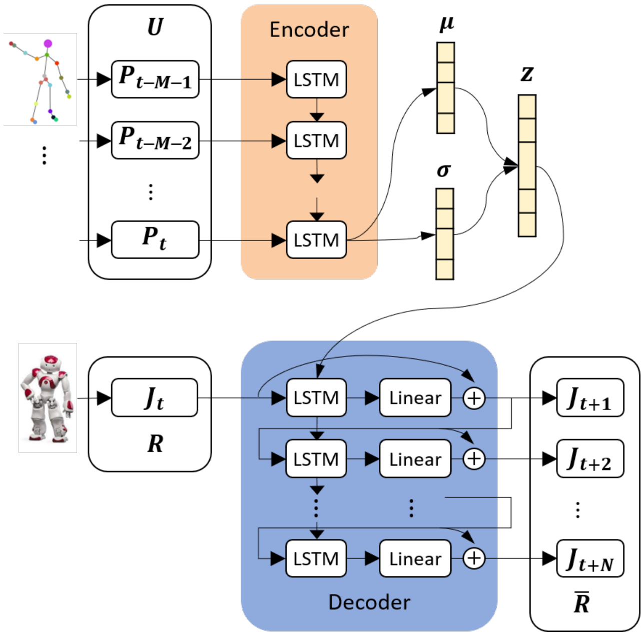


Fig. 1: Overview of our proposed method.

This work was supported by the ICT R&D program of MSIP/IITP. [2017-0-00162, Development of Human-care Robot Technology for Aging Society]

III. Experimental Results

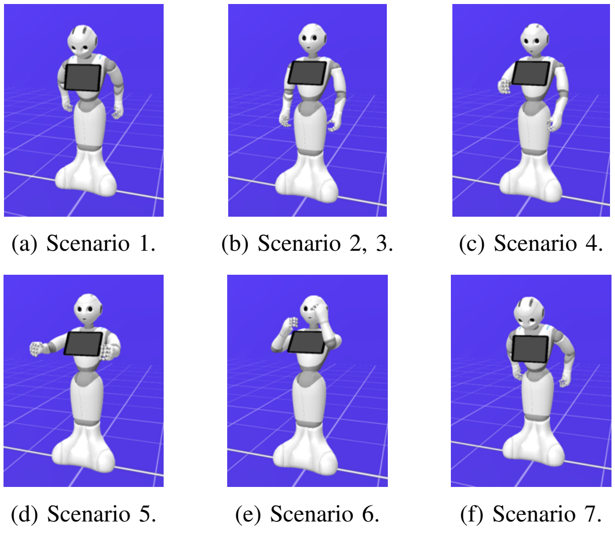
To train and test the model, we used *AIR-Act2Act* dataset [4], which contains 5,000 human-human interaction samples of 10 scenarios. In the experiments, we have selected seven interaction scenarios that occur frequently in daily life: 1) When a person enters into the service area through the door, the other person *bows* to him. 2, 3) When a person walks around (or stands still) without a purpose, the other person *stares* at him. 4) When a person lifts his arm to shake hands, the other person *shakes hands* with him. 5) When a person covers his face and cries, the other person stretches his hands to *hug* him. 6) When a person threatens to hit, the other person *blocks the face* with arms. 7) When a person turns back and walks to the door, the other person *bows* to him.

From each sample, the poses of the person who initiated the interaction were extracted and used as training inputs, i.e. user behaviors, and the poses of the other person were extracted and used as training outputs, i.e. robot behaviors. The numbers of the extracted data are presented in Table I.

TABLE I: The numbers of training and test data.

	Training	Test	Total
Interaction samples	1,575	175	1,750
Extracted data	116,462	12,738	129,200

In the first experiment, we tested the robot behaviors generated in the seven selected interaction scenarios. Figure 2 shows samples of the generated robot behaviors.


(a) Scenario 1. (b) Scenario 2, 3. (c) Scenario 4.

(d) Scenario 5. (e) Scenario 6. (f) Scenario 7.

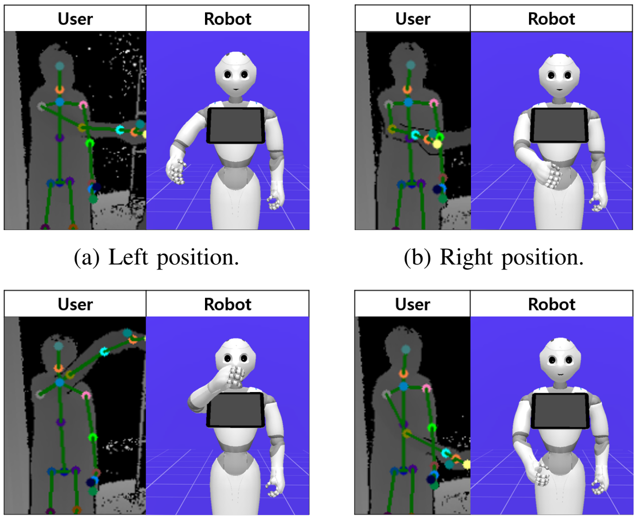
Fig. 2: Robot behaviors generated in 7 interaction scenarios.

Table II shows the success rate of the behavior generation in the seven interaction scenarios. Each interaction scenario has 25 test samples. Experimental results showed that the robot can successfully generate behaviors that correspond to user behaviors.

In the second experiment, we tested the robot behaviors generated when a user lifted his right arm to four different positions to shake hands. Figure 3 illustrates samples of the generated robot behaviors. The experimental results showed that the robot can adjust its behavior according to user postures.

TABLE II: The success rate (SR) of the behavior generation.

Scenario	1	2	3	4	5	6	7	Total
SR [%]	64	80	100	96	96	100	84	88.6



(a) Left position. (b) Right position.

(c) High position. (d) Low position.

Fig. 3: Robot behaviors generated when a user lifted his right arm to different positions.

IV. Conclusions

We propose an end-to-end learning method for generating non-verbal behaviors of social robots. Our model takes the user's previous poses and the current robot pose as inputs and generates the robot's next poses as outputs. Two experiments were carried out using a humanoid robot, Pepper, in a simulated environment. Experimental results showed that the robot can successfully generate multiple social behaviors corresponding to the human behavior, and adjust its behavior according to the user posture.

References

[1] M. Salem, F. Eyssel, K. Rohlfing, S. Kopp, and F. Joublin, "To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability," *International Journal of Social Robotics*, vol. 5, no. 3, pp. 313–323, 2013.

[2] V. Prasad, R. Stock-Homburg, and J. Peters, "Learning human-like hand reaching for human-robot handshaking," *arXiv preprint arXiv:2103.00616*, 2021.

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

[4] W.-R. Ko, M. Jang, J. Lee, and J. Kim, "Air-act2act: Human–human interaction dataset for teaching non-verbal social behaviors to robots," *The International Journal of Robotics Research*, vol. 40, no. 4-5, pp. 691–697, 2021.

[5] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.