International Journal of Social Robotics
https://doi.org/10.1007/s12369-021-00818-1

One-shot Learning from Demonstration Approach Toward a Reciprocal Sign Language-based HRI

Seyed Ramezan Hosseini[1] · Alireza Taheri[1] · Minoo Alemi[1,2] · Ali Meghdari[1]

Accepted: 24 July 2021

Abstract
This paper addresses the lack of proper Learning from Demonstration (LfD) architectures for Sign Language-based Human–Robot Interactions to make them more extensible. The paper proposes and implements a Learning from Demonstration structure for teaching new Iranian Sign Language signs to a teacher assistant social robot, RASA. This LfD architecture utilizes one-shot learning techniques and Convolutional Neural Network to learn to recognize and imitate a sign after seeing its demonstration (using a data glove) just once. Despite using a small, low diversity data set (~500 signs in 16 categories), the recognition module reached a promising 4-way accuracy of 70% on the test data and showed good potential for increasing the extensibility of sign vocabulary in sign language-based human–robot interactions. The expansibility and promising results of the one-shot Learning from Demonstration technique in this study are the main achievements of conducting such machine learning algorithms in social Human–Robot Interaction.

Keywords Human–Robot Interaction (HRI) · Social Robotics · Sign Language · One-shot Learning · Convolutional Neural Network (CNN)

1 Introduction

According to the World Health Organization (WHO), in 2019, over 466 million people (5% of the world's population) had disabling hearing loss [1]. Correspondingly, the 2011 National census reported that about 135,000 Iranians had some kind of hearing or speech difficulties [2]. This sizeable community uses sign language (SL) as one of its most comprehensive ways of communicating [3]. SL is a language utilizing hand gestures and non-manual signs (such as facial expressions) to transfer messages.

A rich and diverse research literature on automatic sign language recognition is to be expected considering the Deaf community's population, cultural diversity, and needs. One of the first works in this field was conducted by Kim et al. in 1996 [4]. They used a pair of Data-Gloves and a fuzzy min–max neural network for the online classification of Korean Sign Language. Liang and Ouhyoung, 1998 [5] concentrated on making sign recognition real-time, continuous, and functional for large vocabularies. This research also utilized Data-Gloves for data acquisition and Hidden Markov Models (HMM) as a recognition algorithm. The authors used features like posture, orientations, and movements to recognize Taiwanese Signs. They reached an 80.4% average recognition rate. The main concern of Vogler and Metaxas' research in 1999 [6] and 2001 [7] was developing scalable solutions for American Sign Language automatic recognition. They referred to Liddell's research on SL phonology and used parallel HMMs to detect each sign. They applied multi-camera processing systems to extract arm movements. Also, in 2001, Kim and Chien [8] used Data-Gloves and HMMs for hand gesture recognition. They decompose gestures into several "strokes" such as right to left or clockwise movements and used them as phonemes to recognize defined gestures. They reported a 96.88% recognition rate. The major contribution of Yang et al.'s research in 2009 [9] was designing a threshold model in a Conditional Random Fields model to increase its recognition rate. They used image processing techniques for data acquisition. In later research in 2009, Cho et al. [10] used semi Markov models for variable-length signs. In a 2016 study, Paudyal

✉ Alireza Taheri
artaheri@sharif.edu

[1] Social and Cognitive Robotics Lab, Sharif University of Technology, Tehran, Iran

[2] Faculty of Humanities, Islamic Azad University, West Tehran Branch, Tehran, Iran

Published online: 10 August 2021

et al. [11] proposed a wearable platform for sign recognition called SCEPTRE. SCEPTRE used two Myo armbands [12] to collect electromyography, gyroscope, and accelerometer data and then utilized Dynamic Time Warping (DTW) algorithms to classify the performed signs. In recent years some researchers used deep neural networks for sign recognition, such as Cui et al. in 2019 [13]. As can be seen, there is no one widely accepted approach in the field, and existing studies have used different algorithms and data collection methods. Also, one of the greatest problems in this field is the extensibility of these algorithms. Most of these works have been tested on only limited sets of signs.

Some researchers have focused on designing SL-based or Gesture-based Human–Robot Interaction (HRI). The most significant researches addressing this issue are as follows: In 2010, Nandy et al. developed an Indian Sign Language based HRI for HOAP-2, an advanced humanoid robotic platform [14]. The system employed several image-processing techniques to extract features from videos of the user's hands. Another Indian Sign Language based HRI was developed in 2017 by Baranwal et al. [15]. This research employed NAO robots as its target platform and used multiple algorithms for sign recognition. After detecting the signed commands, the researchers used the NAO's API and a MATLAB program to engage the user and robot in a conversation with predefined sentences. In 2015, Russo et al. developed a novel telecommunication system for deaf-blind users [16]. The system consists of a robotic hand on one end and a Kinect sensor on the other end. The robotic hand imitates the hand gestures perceived via the Kinect sensor allowing the deaf-blind user to understand the sign by touching it.

Gesture-based HRIs for service robots have a richer research background. One of the first mentions of this subject was in Waldherr et al.'s paper on developing an HRI for a service robot, AMELIA, in 2000 [17]. After implementing their system, the robot was able to detect and obey some gestures common to service robots (Stop, Follow, etc.). Xiao et al.'s notable research in 2014 [18] used a combination of a CyberGlove and Kinect Sensor as a data acquisition setup and various KNN classification algorithms (including Large Margin Nearest Neighbor) to facilitate an upper-body gesture-based interaction between a human user and a humanoid robot, Nadine. The research covered many kinds of interaction, including shaking hands and reacting to the user's actions such as drinking, reading, etc. The extensibility problem of SL detection algorithms is also a common problem in this field, as these proposed HRIs were all created and tested on limited sets of data.

To solve the extensibility problem presented in SL-based HRIs, we believe that such an HRI should also contain the following properties: (1) The recognition module structure must change to a more extensible structure with algorithms capable of learning new signs, and (2) An LfD architecture must be added as a routine to enable the robot to learn new signs from demonstrated signs. To this end, in this paper, we are presenting a solution to improve the human–robot interaction and enhancing the Deaf users' experience by design and implementation of a meaningful gestures learning architecture for robots. A combination of Learning from Demonstration (LfD) and one-shot learning techniques in the architecture's design would enable an SL-based HRI to extend its SL vocabulary (in both recognition and regeneration). This technique would (hopefully) help to develop an appropriate architecture using machine learning algorithms for social Human–Robot Interaction. Unlike the typical situation for applying deep learning algorithms with a large number of training data, the LfD-based algorithm used in this study appropriately works with a limited number of input data.

LfD has shown promising results in aiding the robotic learning process in different areas (such as teleoperation [19–21], rehabilitation [22, 23], robot surgery [24–26], industrial assembly [27], navigation [28–30]) and it has shown suitable effects on the quality of the performance of a social robots' tasks [31, 32]. The LfD design challenge is how to generalize and learn new policies from a small amount of new data. Some good examples of LfD applications can be seen in Calinon et al. researches [33–36], where they try different approaches to teach various robots to recognize and regenerate different kinds of gestures. Another example can be found in [35], where Calinon and Billard use HMMs to teach a robot to recognize and reproduce the English alphabet written by hand movements. Like SL recognition, there is not a commonly accepted algorithm or method in the field of LfD, and completely different algorithms (from probabilistic methods to neural networks) have been used in similar problems. However, in recent years, meta-learning algorithms (like one-shot learning) have become more popular [37–39]. The research done by Finn and colleagues [37] is one of the first examples of using meta-learning techniques in LFD to teach a robot manipulator's various tasks (including reaching, pushing, and placing) in different conditions. Ref. [39] is another more recent example where a robot has been taught to classify objects by seeing only a few examples.

In the following sections, we first become familiar with the robotic platform used in the paper. Then, we design the LfD architecture based on the robot's characteristics and implement it using neural networks. Lastly, we discuss the results and point out some limitations of the study and suggest some points for future work in line with the current study.

2 Methodology

2.1 RASA Robot

Utilizing social robots for children with special needs is increasing in the last decade [40–44]. RASA is a novel social robotic platform whose purpose is to facilitate teaching Iranian Sign Language (ISL) to deaf and hard of hearing Iranian children (Fig. 1). The robot features a cartoon-like face and an attractive exterior. Its interaction modules were designed based on child-robot interaction's requirements. With its 32 Degrees of Freedom (29 DoF in the upper body), its active fingers, and expressive face, RASA can perform comprehensible ISL signs [44, 45]. At the beginning of the current study, RASA had an ISL sign library of about 60 signs.

A customized Cognitive Architecture (CA) has also been developed for RASA [46]. The CA has a highly specialized and modular structure. It has four main modules:

- Perception Unit: accountable for receiving and processing data from the environment.
- Logic Unit: plans for and decides the desired outputs based on the perceived data.
- Action Unit: executes the desired outputs planned by the Logic unit.
- Memory Unit: simultaneously acts as a central junction for data transferring between other units and as a store for structured object-oriented learned data.

In this way, the CA can interact reciprocally with the user through the Perception Unit (as input) and the Action Unit (as output). An overview of the designed CA is shown in Fig. 2 [46]. We used this architecture as a framework for programming the HRI.

2.2 Overview of the LfD Architecture

As is mentioned, RASA is a teaching assistant robot used to help teach ISL to deaf children. It can be employed in a variety of scenarios and interact in many different ways with children or teachers. In this study, we do not restrict the possible HRI experiences, but we assume that in all cases, the robot needs to understand the signs performed for him and answer back in correct distinguishable signs. Our aim is to design an LfD architecture to allow the robot to extend its limited vocabulary.

To teach a new sign to RASA, the teacher needs to wear the data glove and perform the sign (at least once) to pass the corresponding word to the robot. A pre-trained neural network converts the sign to an embedded vector. New signs will be recognized by comparing them to these embedded vectors. This is the learning process for the sign recognition portion. The sign imitation is done by mapping the performed sign to the robot's kinematics in a way similar to [47] (see Fig. 3).

We use one-shot learning techniques in pre-training the neural network. The network's output is an embedded vector and not categorical in order to enable it to perform in new categories outside of the training output. In the training process, we train this network on a diverse dataset to expose it to many kinds of features and force it (with the aid of a loss function) to map the input data in clusters with sufficient margins. In this way, the network can map new signs in new vectors, and hopefully, they will form new clusters. Then,

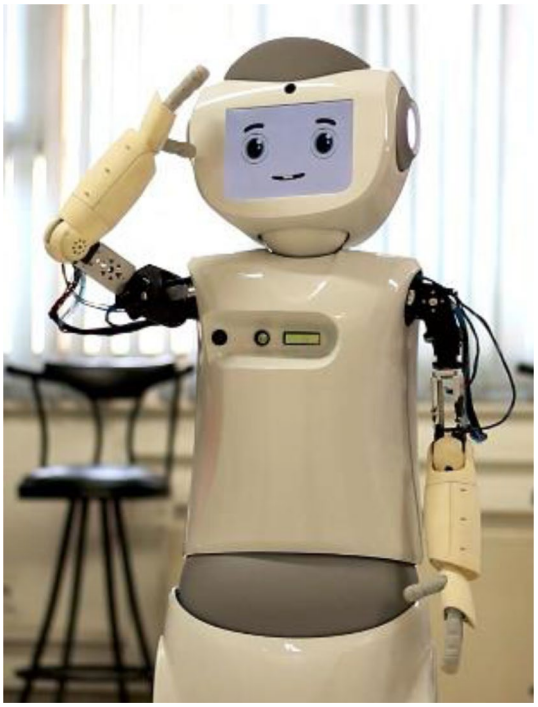

Fig. 1 The RASA Robot

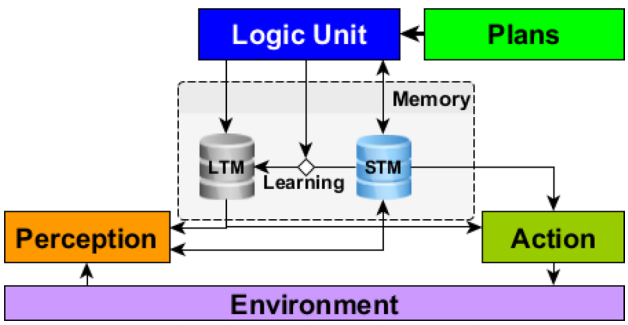

Fig. 2 The general overview of RASA's cognitive architecture [46]

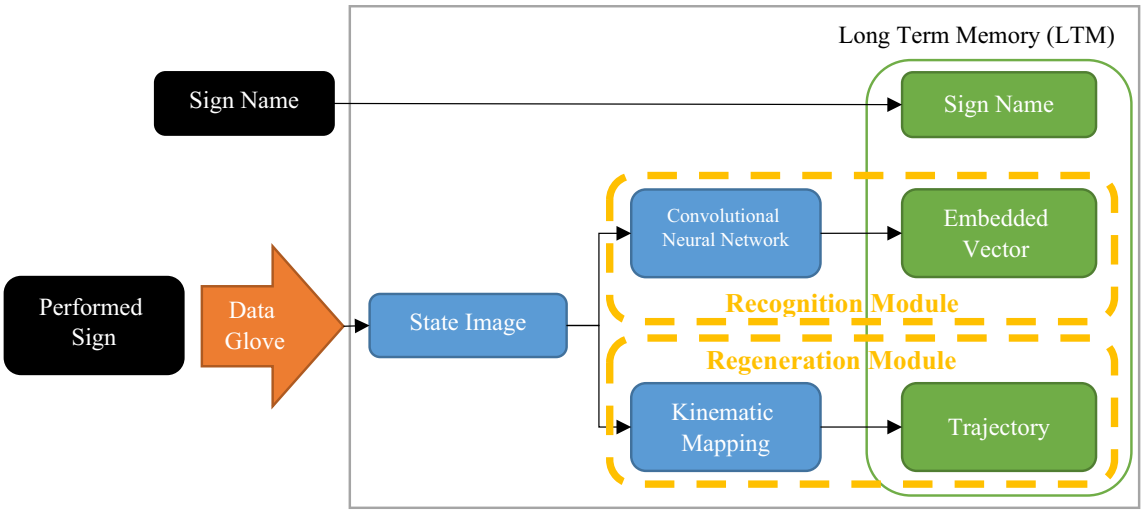

Fig. 3 An overview of the proposed LfD architecture in this study. A user gives a sign's name to the robot and performs that sign using a Data Glove. The architecture converts the performed sign to a state image and feeds it to its recognition and regeneration modules. The recognition module feeds the incoming state image to a pre-trained convolutional neural network and stores its vector output alongside the given name in the robot's Long Term Memory (see Fig. 2). The regeneration module maps the given state image to the robot's kinematics and stores the output trajectory in the memory. This is the learning process. Now the robot can retrieve the trajectory or the embedded vector for recognizing or regenerating that sign in the future

the classification can be done by other criteria such as the Nearest Neighbor.

2.3 Implementation of the LfD Architecture

2.3.1 Data Collection

We chose a Neuron Lite glove from Noitom Ltd. for this study [48]. It is a single-arm sensor glove using six sensors, each with a 3-axis accelerometer, 3-axis gyroscope sensor, and 3-axis magnetometer. The sensors are located on the middle of the arm, on the wrist, on the back of the hand, and on the tip of the thumb, index, and middle fingers of the right hand. So it can detect the movements of the right hand including its three fingers. The outputs of the glove are in the form of joint positions and rotations. The glove connects to a proprietary software called "Axis Neuron" [49]. Axis Neuron can save motions (in various file formats) for offline use. In addition, it can use various communication protocols to broadcast motion data (in BVH format) in real-time.

In the next step, we chose 16 ISL signs. All of the selected signs were distinguishable from features that could be extracted by the 3-finger data glove. We also preferred to have different categories of signs: static signs, dynamic signs with simple motions, and signs with periodic motions. The selected signs are comprised of several color signs and common signs, including yes/no and greetings (Fig. 4).

Due to COVID-19 limitations, we have gathered a narrow dataset. After teaching 12 hearing adults (unfamiliar with sign language) the signs, we asked them to perform selected signs. Each sign was performed 3 times in different order. We recorded these performances with the Axis Neuron and saved them to a dataset as ".BVH" files. After removing bad demonstrations, each sign has more than 30 performances

Fig. 4 Iranian Signs selected for the HRI [3]

and thus the final dataset had approximately 500 signs in 16 categories.

2.3.2 State Images

The recorded BVH files contain full data for every joint in the participants' body for the whole performance duration. As the glove only gathers data from the right arm and the performers paused many times between the signs, a good deal of the files' contents are unnecessary. Therefore, a necessary step is trimming, segmenting files, and feature selection.

For feature selection, we refer to Stokoe's theory. In Stokoe's theory [50, 51], every (manual) sign consists of four basic elements: Hand Shape, Palm Orientation, Hand Location, and Movements (changes in the first three elements). Therefore, unlike a spoken language, the phonemes of signs in SL are occurring simultaneously. Since there are not enough studies on ISL structure, the current study assumes that Stokoe's theories are extensible to ISL. Amongst all the possible features of these elements, we chose:

- Hand location (a height-normalized 3d vector from the shoulder toward the hand in the Cartesian coordination).
- Palm orientation (as a quaternion).
- Handshape (normalized angles representing the fingers' flexion).

In the next step, we have scaled and resampled all demonstration times so we could make a 12×50 matrix representing each demonstration. These image-like matrixes are called State Images [52, 53]. In Fig. 5, readers can see their structure and a sample State Image of the sign "Orange". These images are the inputs of our models.

2.3.3 Network Architecture

To classify the performed signs, we used a simple Convolutional Neural Network to map each State Image to a 512-dimensional vector. The proposed structure is shown in Fig. 6.

The chosen structure was kept as simple as possible to allow us to train it with our limited dataset of ~ 500 signs in the meta-learning phase. As our purpose is to have an expansible vocabulary, the network also maps the state images to an embedding vector instead of returning a categorical result or one-hot.

The proposed architecture has 2,462,912 trainable parameters, which is very large for our limited dataset. We chose the Siamese networks [56] with cosine triplet loss [57] to

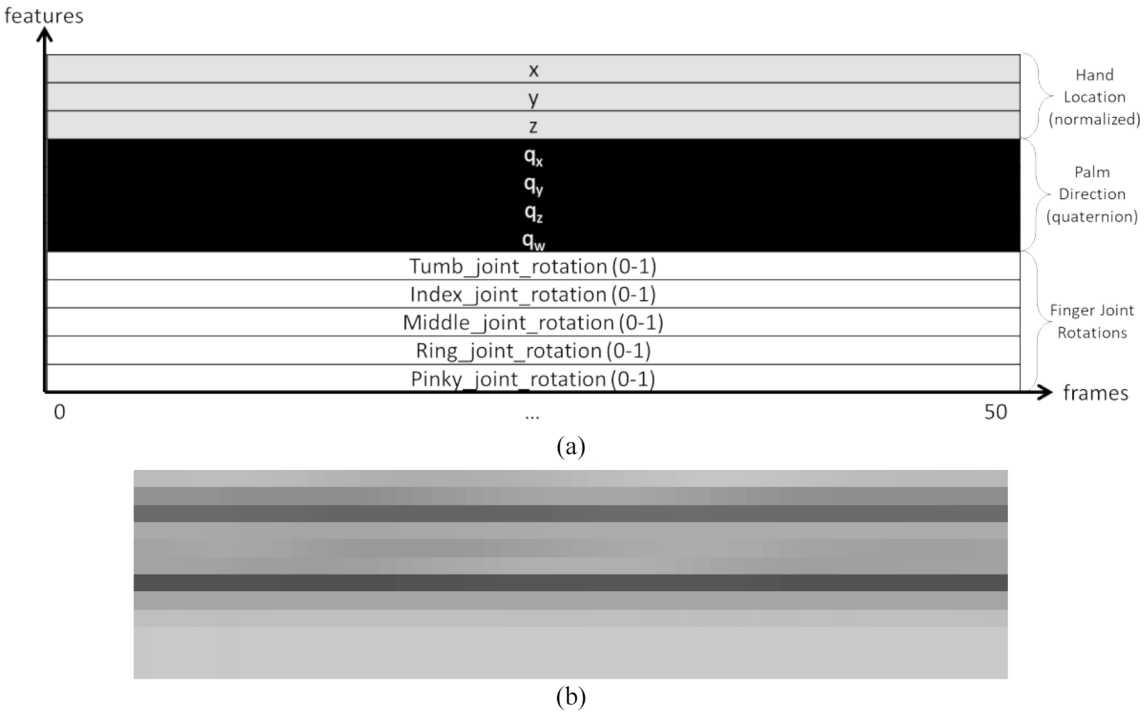

Fig. 5 State Image. **a** The structure of the State Images. **b** State Image sample of the sign "Orange" [54]. In the "Orange" sign the handshape and palm direction are fixed so the last 9 rows of the image do not show significant changes in their values over time. The thumb and index are closed so the 7th and 8th rows (from top) are slightly darker than the rows beneath them. The hand moves in a circle in the x–y plane so the sinusoidal movement in the 1st and 2nd rows are clear, but there is no change in the 3rd row (associated with the z-axis)

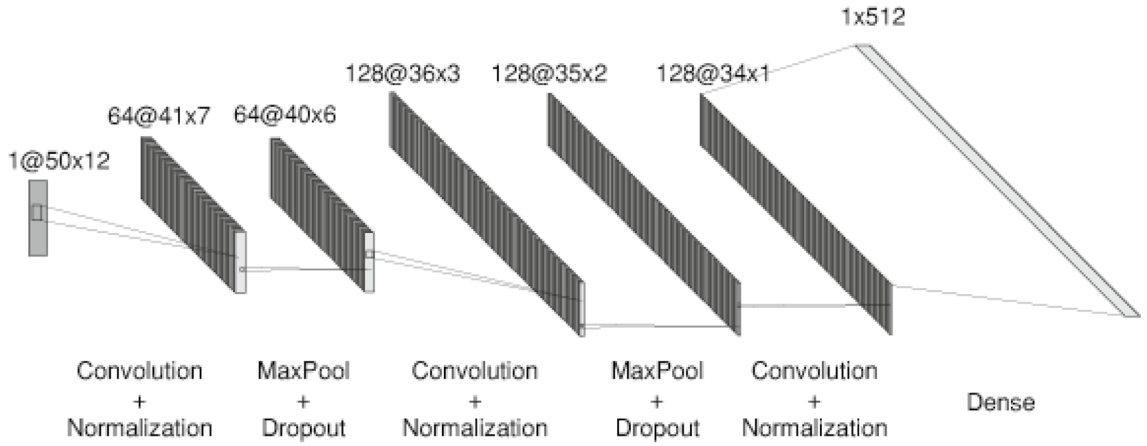

Fig. 6 The chosen CNN Architecture [55]

train the CNN with the smaller amount of data from this dataset. In this architecture, we have three parallel CNNs with shared weights. We feed three images simultaneously to them and compute the triplet loss of the vectors produced by them. The first image is called the anchor image. The second image (also known as the positive images) must be an image with the same label, and the third image (also known as the negative image) must belong to a different class. Using this method, we have more than eight million data entries.

The loss function is as follows (Eq. 1):

$$loss(a, p, n) = \max(0, d(a, p) - d(a, n) + \alpha) \tag{1}$$

where $d(x, y)$ is the distance of x and y (predicted vectors of images X and Y, respectively) and is calculated as (Eq. 2):

$$d(x, y) = 1 - \hat{x}.\hat{y}, \hat{x} = \frac{x}{|x|} \tag{2}$$

This loss function tries to maximize the distance between anchor images and negative images while keeping anchors and positives close enough together. α is a margin parameter to eliminate the chance of converging to a trivial answer (mapping all images to zero vectors). Following the results in [58], we chose $\alpha = 0.2$.

2.3.4 Training

The designed architecture was implemented using the Keras library [59]. We used stochastic gradient descent (without momentum and with a learning rate of 0.001) as the optimizer. Google Colab platform [60] is used to train the model (batch size = 128). Moreover, we used the following methods to improve the performance and prevent overfitting issues:

- Reducing the learning rate by 80 percent after 4 steps of no improvements on validation loss;
- Implementing early stopping with the patience of 10 steps on validation loss; and
- Feeding negative images based on their current distance to the anchor image (nearest to the anchor image).

To check the extensibility of the model, we trained the model 16 times, and each time we exclude one class from the training process to be used as test data (~ 6% of all of the data). We also separated ~ 20% of the remaining data (equally distributed in 15 remaining classes) as the validation data.

2.3.5 Evaluation

We check the model's accuracy using n-way accuracy ($n = 4$, 8, 12, and 16). In this method, we repeat a testing procedure 1000 times. In each round, we randomly select an anchor image within the target data and then randomly draw n other images as benchmark images. These benchmark images consist of one positive image (an image with the same label as the anchor image) and n-1 negative images (images with different labels than the anchor image). We calculate the similarity between the anchor image and benchmark images in terms of $d(x, y)$. We consider this round a "True" case if the nearest benchmark image to the anchor image is a positive image; otherwise, it is considered a "False" case. After completing the procedure, the accuracy was computed by dividing the number of "True" cases by the number of rounds (i.e., 1000). To better understand the model's extensibility, we perform this test within two groups: first, the validation data within the classes present in the training data, and second, the test data (from the excluded class). The test is performed four times for each group in all 16 training sessions.

3 Results and Discussions

The main result of the implementation of the proposed structure is an LfD plugin that enables the robot to learn new signs just by watching it once. We can see the performance of the one-shot sign recognition module in Table 1. The full results are presented in Table 2 in Appendix.

It can be seen that as a general trend, the accuracy in both trial modes decreases when the n in the n-way procedure increases, which is expected but may raise concerns about the performance after the expansion of the signs library. There was no significant difference ($p > 0.05$ in all cases) between the mean accuracies in the test and training trials, which is quite appropriate. Although in terms of the Standard Deviations (SD), we observed some meaningful differences. In general, the SD of the training data trials is less, which shows the robustness of the architecture in the absence of some data in the meta-learning phase. Trials of the test data show more SD, which is expected because the module is testing against signs it has not seen before. Hence, some signs are less likely to be recognized by the module (such as Pink and Purple), and some (such as Black and Orange) are more recognizable by the module. We assume this is because of some unique features in the more recognizable signs (i.e. unique movements of hands in Black and unique hand shape in Orange) and the lack of similarity of the less recognizable signs to the training signs.

Table 1 Summary of the recognition module's accuracy (percentage) including the mean and the standard deviations of all the trials (For more details, see Table 2 in the Appendix)

16-way	12-way	8-way	4-way	
45.89 ± 9.84	50.54 ± 9.69	57.81 ± 9.16	70.83 ± 8.02	Test
46.75 ± 2.88	51.55 ± 2.93	58.25 ± 2.33	71.66 ± 2.06	Train

While using one-shot learning strategies in LfD is not unheard of [61, 62], meaningful gesture learning research is so uncommon that we have yet to come across any. Therefore, due to the lack of similar papers in this field, it is somewhat impossible for us to do a thorough comparison between the findings of this study and other related works. Nevertheless, the expansibility and promising results of the one-shot Learning from Demonstration technique are in line with other researches [61, 62] which are the main achievements of conducting such machine learning algorithms in social Human–Robot Interaction. In terms of accuracy, the implemented system is weaker than many of state of the art researches [63–66]; however, due to the small size of the dataset and other limitations, this is justifiable and promising. It should be noted that the same level of accuracy is reported in recent papers in the literature (such as [13]).

4 Limitations and Future Work

This study had serious limitations (partly because of the COVID-19 pandemic) in the dataset size (less than 600 demonstrations), diversity (just 16 signs), and gathering process (lack of deaf community presence and limited data glove). Despite these limitations, the results show promise that with a richer and more diverse base dataset, we can reach an accurate and extensible LfD architecture. Also, we wish we could have conducted in vivo HRI experiments to investigate the users' opinions about this framework and their experience while using it. But unfortunately, due to COVID-19 restrictions, it was not possible for us to conduct experiments.

The ultimate goal of the current research line is to enable RASA to act as a teaching assistant robot in SL teaching environments. The present study followed this line; therefore, we suggest future works along the subsequent line in order to make RASA faultless in learning and teaching new signs and interacting with people using its knowledge in SL. Keeping these main goals in mind, the recommended directions in extending the current study lies in the following routes:

- Removing the limitations: focusing on making a richer, more diverse dataset with more signs and more demonstrations, more demonstrators familiar with sign language, and better equipment to record more com-

plex signs; conducting HRI based trials to measure the acceptability, ease of use, and desirability of the results.
- Enhancing the recognition module: trying to make the module more accurate or more extensible by using better data or trying other algorithms and techniques (such as data augmentation).
- Enhancing the imitation module: implementing methods with more natural results than just mimicking the movements (such as GMM/GMR approaches in [33–36], Variational Auto-Encoders (VAE) or Generative Adversarial Networks (GAN)).

5 Conclusion

Using one-shot learning strategies and Convolutional Neural Networks, we introduced and implemented an LfD-based system to teach new ISL signs to a teacher assistant robot in this research. The proposed architecture can learn to recognize and imitate a sign after just one demonstration using a data glove. The reorganization module reached a 4-way accuracy of 70% on the test data, and while this is not a very high accuracy level, it was still promising if we consider the small size and low diversity of the data set used in this study. The module showed good potential to make the HRIs more extensible. The main point is that the presented results have been obtained from a small number of training data in contrast to the typical deep learning algorithms which need big datasets. The expansibility and promising results of the one-shot Learning from Demonstration technique are the main achievements of conducting such machine learning algorithms in social Human–Robot Interaction.

Appendix

See Table 2.

Table 2 Recognition module's accuracy (percentage) in Detail for all the trials. (Color figure online)

Train				Test				Test Class
16-way	12-way	8-way	4-way	16-way	12-way	8-way	4-way	
47.7	50.5	55.0	71.9	49.9	55.2	62.2	75.8	Word #1: "Hi"
48.8	51.9	58.1	69.6	48.6	53.1	63.0	76.6	
50.2	54.0	58.4	70.7	49.7	52.8	62.6	74.6	
46.9	56.1	60.3	71.5	48.2	51.9	61.2	73.7	
52.2	53.5	57.4	70.1	45.1	48.7	58.8	76.7	Word #2: "Bye"
50.3	53.7	58.6	70.7	47.2	47.2	59.7	76.4	
46.3	48.7	59.1	72.7	46.4	47.0	59.7	76.6	
48.9	54.6	59.2	71.7	43.4	50.3	60.1	75.8	
50.3	52.0	57.7	72.8	44.2	50.8	58.8	70.2	Word #3: "True"
49.9	56.2	61.6	73.9	45.9	49.4	56.0	69.8	
47.4	52.8	60.7	74.5	45.2	46.6	57.7	66.2	
50.2	53.8	60.2	75.9	43.4	48.5	55.4	70.2	
45.9	51.7	59.5	71.9	46.0	49.6	61.1	71.9	Word #4: "False"
47.2	51.6	57.8	72.8	46.9	52.9	58.2	73.1	
45.9	51.2	57.8	72.5	44.4	53.2	59.9	73.3	
45.7	50.4	56.4	72.2	45.5	50.3	60.7	75.9	
44.0	44.8	56.3	70.2	41.6	45.7	52.4	67.4	Word #5: "Yes"
45.3	50.3	56.1	70.5	40.5	46.3	51.0	66.4	
41.8	47.6	55.7	70.2	44.1	46.2	53.8	67.0	
44.6	48.2	55.3	72.6	43.2	46.6	54.4	67.2	
44.8	49.8	59.9	71.7	43.4	45.4	56.4	69.6	Word #6: "No"
45.5	53.1	57.4	74.2	38.7	48.3	53.9	67.3	
47.0	52.5	58.7	70.7	40.8	46.6	56.3	68.8	
45.4	49.4	57.9	68.5	42.6	44.4	50.9	70.9	
47.0	49.7	55.5	69.0	51.3	53.5	61.2	71.0	Word #7: "Resign"
45.0	47.5	58.1	68.7	50.5	56.2	60.9	72.1	
46.7	50.0	56.4	70.8	50.5	53.4	60.1	68.0	
46.6	50.4	60.1	69.4	48.5	56.8	60.0	73.9	
46.4	52.3	55.9	72.1	37.1	40.2	49.8	66.7	Word #8: "Back"
43.2	52.8	58.0	72.0	35.6	43.0	52.3	68.1	
46.0	52.7	56.0	68.9	35.5	43.0	50.6	66.2	
48.3	53.7	59.4	71.0	37.2	43.3	49.0	66.0	
41.1	46.4	54.9	69.0	55.2	60.3	68.3	80.4	Word #9: "Black"

Table 2 (continued)

41.2	45.2	54.0	67.0	59.1	60.5	67.3	80.8	
41.2	48.6	58.6	68.5	56.4	63.6	68.8	78.9	
39.7	46.9	54.9	68.5	56.5	61.0	66.4	79.8	
50.9	58.1	64.0	76.6	44.4	47.2	53.4	66.0	
50.7	56.3	64.4	73.7	42.7	48.7	54.4	66.9	Word #10: “Red”
53.0	57.9	64.7	75.8	44.0	47.6	55.0	67.0	
54.0	57.0	61.5	75.5	45.9	46.4	52.9	65.6	
44.9	52.0	57.5	70.6	46.1	58.0	60.2	72.0	
46.7	50.3	59.8	70.5	49.0	54.2	60.6	72.2	Word #11: “Green”
47.9	53.4	59.5	70.1	47.8	54.1	61.9	74.9	
47.9	50.3	60.0	72.7	46.2	50.0	58.0	72.8	
44.5	48.2	56.8	69.0	52.7	60.4	67.1	76.8	
44.1	49.6	55.9	68.7	52.8	59.9	65.8	78.1	Word #12: “Brown”
44.1	47.4	56.3	70.4	55.1	59.5	67.3	78.4	
45.5	47.6	54.5	69.2	53.5	60.5	65.0	77.0	
48.0	52.8	62.9	74.2	43.0	46.2	54.9	66.4	
48.5	55.2	59.6	72.9	40.8	46.9	54.3	64.3	Word #13: “Blue”
46.0	54.8	60.6	74.4	44.9	46.6	52.5	67.0	
47.0	54.4	58.9	72.4	40.4	47.0	51.0	64.5	
45.5	53.4	59.6	72.4	24.7	33.1	35.7	52.8	
48.6	52.7	59.7	72.3	25.7	29.5	39.2	48.8	Word #14: “Pink”
48.6	51.5	57.7	73.9	26.5	29.8	38.7	50.7	
51.0	50.5	57.9	72.4	27.6	34.5	37.9	50.2	
47.5	52.2	56.4	73.2	73.4	75.3	80.0	87.6	
49.2	52.9	57.1	72.0	69.2	76.4	81.0	89.9	Word #15:
49.1	50.1	60.1	72.1	71.9	76.0	79.8	87.8	“Orange”
43.7	48.6	57.7	73.4	74.4	74.3	79.9	84.8	
46.2	51.2	56.8	70.8	37.7	39.2	46.8	64.0	
44.0	54.2	56.1	71.5	36.1	38.5	48.7	65.9	Word #16: “Purple”
44.0	50.0	56.3	72.3	37.0	42.4	49.8	64.8	
46.3	52.3	59.1	74.1	35.2	40.7	49.0	62.8	

Acknowledgements This research was supported in part by the "Dr. Ali Akbar Siassi Memorial Research Grant Award" and the "Iranian National Science Foundation (INSF)" (http://en.insf.org/) (Grant No. 98025100).

Declarations

Conflict of interest Author Alireza Taheri has received research a grant from the "Iranian National Science Foundation (INSF)" (Grant No. 98025100). The authors Seyed Ramezan Hosseini, Ali Meghdari, and Minoo Alemi declare that they have no conflict of interest.

References

1. WHO factsheet on deafness and hearing loss (2019) World Health Organization. Tech Rep Mar [Online]. Available: http://www.who.int/mediacentre/factsheets/fs300/en/
2. Iran 2011 census full report (2011) Statistical Centre of Iran, Tech Rep
3. Persian Sign Language (2010) 4th ed., Sign Language Research Committee, University of Social Welfare and Rehabilitation Sciences, Tehran, Iran
4. Kim J-S, Jang W, Bien ZZ (1996) A dynamic gesture recognition system for the Korean sign language (KSL). IEEE Trans Syst Man Cybern B Cybern 26(2):354–359. https://doi.org/10.1109/3477.485888
5. R.-H. Liang and M. Ouhyoung, (1998) A real-time continuous gesture recognition system for sign language. In: proceedings third IEEE international conference on automatic face and gesture recognition. Nara, Japan, pp. 558–567. DOI: https://doi.org/10.1109/AFGR.1998.671007
6. C. Vogler and D. Metaxas, (1999) Parallel hidden Markov models for American sign language recognition. In: ICCV, Kerkyra, Greece, vol. 1, pp. 116-122. DOI: https://doi.org/10.1109/ICCV.1999.791206
7. Vogler C, Metaxas DN (2001) A framework for recognizing the simultaneous aspects of American sign language. Comput Vis Image Underst 81(3):358–384. https://doi.org/10.1006/cviu.2000.0895
8. Kim I-C, Chien S-I (2001) Analysis of 3D hand trajectory gestures using stroke-based composite hidden Markov models. Appl Intell 15(2):131–143. https://doi.org/10.1023/A:1011231305559
9. Yang H-D, Sclaroff S, Lee S-W (2009) Sign language spotting with a threshold model based on conditional random fields. IEEE Trans Pattern Anal Mach Intell 31(7):1264–1277. https://doi.org/10.1109/TPAMI.2008.172
10. S.-S. Cho, H.-D. Yang, and S.-W. Lee, (2009) Sign language spotting based on semi-Markov conditional random field. In: WACV, Snowbird, Utah, USA, DOI: https://doi.org/10.1109/WACV.2009.5403109
11. P. Paudyal, A. Banerjee, and S. K. Gupta, (2016) SCEPTRE: a pervasive, non-invasive, and programmable gesture recognition technology. In: IUI 16, Sonoma, California, USA, DOI: https://doi.org/10.1145/2856767.2856794
12. Myo gesture control armband|wearable technology by thalmic labs, Accessed: (2018) [Online]. Available: https://www.myo.com/
13. Cui R, Liu H, Zhang C (2019) A deep neural framework for continuous sign language recognition by iterative training. IEEE Trans Multimed 21(7):1880–1891. https://doi.org/10.1109/tmm.2018.2889563
14. A. Nandy, S. Mondal, J. S. Prasad, P. Chakraborty and G. C. Nandi, (2010) Recognizing & interpreting Indian Sign Language gesture for Human Robot Interaction. In: ICCCT, Allahabad, Uttar Pradesh, India, pp. 712–717. DOI: https://doi.org/10.1109/ICCCT.2010.5640434
15. Baranwal N, Singh AK, Nandi GC (2017) Development of a framework for human-robot interactions with Indian sign language using possibility theory. Int J of Soc Robotics 9(4):563–574. https://doi.org/10.1007/s12369-017-0412-0
16. Russo LO, Farulla GA, Pianu D, Salgarella AR, Controzzi M, Cipriani C, Oddo CM, Geraci C, Rosa S, Indaco M (2015) PARLOMA; a novel human-robot interaction system for deaf-blind remote communication. Int J Adv Robot Syst 12(5):57. https://doi.org/10.5772/60416
17. Waldherr S, Romero R, Thrun S (2000) A gesture based interface for human-robot interaction. Auton Robots 9(2):151–173. https://doi.org/10.1023/A:1008918401478
18. Xiao Y, Zhang Z, Beck A, Yuan J, Thalmann D (2014) Human-robot interaction by understanding upper body gestures. Presence-Teleop Virt 23(2):133–154. https://doi.org/10.1162/PRES_a_00176
19. Havoutis I, Calinon S (2017) Supervisory teleoperation with online learning and optimal control. In: 2017 IEEE International Conference on Robotics and Automation (ICRA) pp. 1534–1540
20. Birk A, Doernbach T, Mueller C, Luczynski T, Chavez AG et al (2018) Dexterous underwater manipulation from onshore locations: streamlining efficiencies for remotely operated underwater vehicles. IEEE Robot Autom Mag 25:24–33
21. Su H, Qi W, Yang C, Sandoval J, Ferrigno G, De Momi E (2020) Deep neural network approach in robot tool dynamics identification for bilateral teleoperation. IEEE Robot Autom Lett 5(2):2943–2949
22. Lauretti C, Cordella F, Guglielmelli E, Zollo L (2017) Learning by demonstration for planning activitiesof daily living in rehabilitation and assistive robotics. IEEE Robotics and Automation Letters 2:1375–1382
23. Fong J, Tavakoli M (2018) Kinesthetic teaching of a therapist's behavior to a rehabilitation robot. In: 2018 International Symposium on Medical Robotics (ISMR). pp. 1–26
24. Najafi M, Sharifi M, Adams K, Tavakoli M (2017) Robotic assistance for children with cerebral palsybased on learning from tele-cooperative demonstration. Int J Intell Robot Appl 1:43–54
25. Su H, Qi W, Hu Y, Karimi HR, Ferrigno G, De Momi E (2020) An incremental learning framework for human-like redundancy optimization of anthropomorphic manipulators. IEEE Transactions on Industrial Informatics
26. Su H, Mariani A, Ovur SE, Menciassi A, Ferrigno G, De Momi E (2021) Toward teaching by demonstration for robot-assisted minimally invasive surgery. IEEE Trans Autom Sci Eng 18(2):484–494
27. Zhu Z, Hu H (2018) Robot learning from demonstration in robotic assembly: a survey. Robotics 7:17
28. Pomerleau DA (1991) Efficient training of artificial neural networks for autonomous navigation. Neural Comput 3:88–97
29. Boularias A, Kromer O, Peters J (2012) Structured apprenticeship learning. In: joint European conference on machine learning and knowledge discovery in databases. Springer
30. Pan Y, Cheng CA, Saigol K, Lee K, Yan X, et al (2018) Agile autonomous driving using end-to-end deep imitation learning. In: robotics: science and systems
31. Hosseini SR, Taheri A, Meghdari A, & Alemi M (2019). Teaching persian sign language to a social robot via the learning from demonstrations approach. In: international conference on social robotics (pp. 655–665). Springer, Cham
32. Fard AR, Hosseini SR, Taheri A, & Meghdari A (2020) Can learning from demonstration reproduce natural and understandable

movements?, In: 2020 8th international conference on robotics and mechatronics (ICRoM), Tehran, Iran
33. Rozo L, Silvério J, Calinon S, Caldwell DG (2016) Learning controllers for reactive and proactive behaviors in human-robot collaboration. Front Robot AI 3(30):1–11. https://doi.org/10.3389/frobt.2016.00030
34. S Calinon and A Billard, (2004) Stochastic gesture production and recognition model for a humanoid robot. In: IROS, Sendai, Japan, vol. 3, pp. 2769–2774, DOI: https://doi.org/10.1109/IROS.2004.1389828
35. S Calinon and A Billard, (2005) Recognition and reproduction of gestures using a probabilistic framework combining PCA, ICA and HMM. In: ICML 05, Bonn, Germany, pp. 105–112. DOI: -https://doi.org/10.1145/1102351.1102365
36. Calinon S (2016) A tutorial on task-parameterized movement learning and retrieval. Intel Serv Robot 9(1):1–29. https://doi.org/10.1007/s11370-015-0187-9
37. Finn C, Yu T, Zhang T, Abbeel P, Levine S (2017) One-shot visual imitation learning via meta-learning. arXiv preprint
38. Lu J, Gong P, Ye J, Zhang C (2020) Learning from Very Few Samples: A Survey. arXiv preprint
39. Ayub A, Wagner AR (2020) Tell me what this is: few-shot incremental object learning by a robot. arXiv preprint
40. Pour AG, Taheri A, Alemi M, Meghdari A (2018) Human–robot facial expression reciprocal interaction platform: case studies on children with autism. Int J Soc Robot 10(2):179–198
41. Alemi M, Taheri A, Shariati A, Meghdari A (2020) Social robotics, education, and religion in the Islamic World: an Iranian perspective. J Sci Eng Ethics 26(3):1–26
42. Taheri A, Meghdari A, Alemi M, Pouretemad H (2018) Human–robot interaction in autism treatment: a case study on three pairs of autistic children as twins, siblings, and classmates. Int J Soc Robot 10(1):93–113
43. Shahab M, Taheri A, Hosseini SR, Mokhtari M, Meghdari A, Alemi M, & Pour AG (2017). Social Virtual reality robot (V2R): a novel concept for education and rehabilitation of children with autism. In: 2017 5th RSI international conference on robotics and mechatronics (ICRoM) (pp. 82–87). IEEE
44. M Zakipour, A Meghdari, and M Alemi, (2016) Rasa: a low-cost upper-torso social robot acting as a sign language teaching assistant. In: ICSR, Kansas City, USA, pp. 630–639. DOI: https://doi.org/10.1007/978-3-319-47437-3_62
45. Meghdari A, Alemi M, Zakipour M, Kashanian SA (2018) Design and realization of a sign language educational humanoid robot. J Intell Robot Syst. https://doi.org/10.1007/s10846-018-0860-2
46. SR Hosseini, A Taheri, A Meghdari, and M Alemi, (2018) "Let there be intelligence!"- a novel cognitive architecture for teaching assistant social robots. In: ICSR, Qingdao, China, pp. 275–285. DOI: https://doi.org/10.1007/978-3-030-05204-1_27
47. Taheri A, Meghdari A, Mahoor MH (2020) A close look at the imitation performance of children with autism and typically developing children using a robotic system. Int J Soc Robot. https://doi.org/10.1007/s12369-020-00704-2
48. Neuron lite glove specifications, (2017) Accessed. [Online]. Available: https://web.archive.org/web/20170703221605/https://neuronmocap.com/content/product/perception-neuron-lite
49. Axis neuron software home page, (2018) Accessed. [Online]. Available: https://web.archive.org/web/20181002123022/https://neuronmocap.com/content/axis-neuron-software
50. Stokoe WC (2005) Sign language structure: an outline of the visual communication systems of the American deaf. J Deaf Stud Deaf Educ 10(1):3–37. https://doi.org/10.1093/deafed/eni001
51. Stokoe WC (1980) Sign language structure. Annu Rev Anthropol 9(1):365–390. https://doi.org/10.1146/annurev.an.09.100180.002053
52. Dong Y (2018) An application of deep neural networks to the in-flight parameter identification for detection and characterization of aircraft icing. Aerosp Sci Technol 77:34–49
53. Dong Y (2019) Implementing deep learning for comprehensive aircraft icing and actuator/sensor fault detection/identification. Eng Appl Artif Intell 83:28–44
54. Hunter JD (2007) Matplotlib: A 2D graphics environment. Comput Sci Eng 9(3):90–95
55. LeNail A (2019) NN-SVG: publication-ready neural network architecture schematics. J Open Source Soft 4(33):747. https://doi.org/10.21105/joss.00747
56. Bromley J, Bentz JW, Bottou L, Guyon I, LeCun Y, Moore C et al (1993) Signature verification using a "siamese" time delay neural network. Int J Pattern Recognit Artif Intell 7(04):669–688
57. Novoselov S, Shchemelinin V, Shulipa A, Kozlov A, & Kremnev I (2018). Triplet loss based cosine similarity metric learning for text-independent speaker recognition. In: Interspeech (pp. 2242–2246)
58. Schroff F, Kalenichenko D, & Philbin J (2015). Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 815–823)
59. Chollet, F. et al (2015) Keras. https://keras.io.Accessed 23 May 2021
60. Bisong E (2019) Google Colaboratory. In: building machine learning and deep learning models on google cloud platform. Apress, Berkeley, CA. doi: https://doi.org/10.1007/978-1-4842-4470-8_7
61. Duan Y, Andrychowicz M, Stadie B, Ho OJ, Schneider J, Sutskever I, & Zaremba W (2017) One-shot imitation learning. In: Advances in neural information processing systems (pp. 1087–1098)
62. Lafleche JF, Saunderson S, Nejat G (2018) Robot cooperative behavior learning using single-shot learning from demonstration and parallel hidden Markov models. IEEE Robot Automat Lett 4(2):193–200
63. M Taskiran, M Killioglu, and N Kahraman, (2018) A real-time system for recognition of American sign language by using deep learning. In: 2018 41st international conference on telecommunications and signal processing (TSP), Athens, Greece, IEEE, doi: https://doi.org/10.1109/TSP.2018.8441304
64. Tang A, Lu K, Wang Y, Huang J, Li H (2015) A real-time hand posture recognition system using deep neural networks. ACM Trans Intell Syst Technol 6(2):1–23. https://doi.org/10.1145/2735952
65. Meghdari A, Alemi M (2018) Recent advances in social & cognitive robotics and imminent ethical challenges. In: Proceedings of the 10th international RAIS conference on social sciences and humanities, Vol. 211, pp. 75–82
66. Zibafar A, Saffari E, Alemi M, Meghdari A, Faryan L, Pour AG, Taheri A (2019) State-of-the-art visual merchandising using a fashionable social robot: RoMa. Int J Soc Robot 13(3):509–523

Seyed Ramezan Hosseini is a PhD candidate at Mechanical Engineering Department of Sharif University of Technology, Tehran, Iran. His research interests include Social Robotics, Artificial Intelligence, Machine Learning, and their applications in Iranian Sign Language.

Alireza Taheri is an Assistant Professor of Mechanical Engineering with an emphasis on Social and Cognitive Robotics at Sharif University of Technology, Tehran, Iran. He is the Head of the Social and Cognitive Robotics Lab at Sharif University of Technology.

Minoo Alemi received her Ph.D. in Applied Linguistics from Allameh Tabataba'i University in 2011. She is currently an Associate Professor and Division Head of Applied Linguistics at the Islamic Azad University, West-Tehran Branch. She is the co-founder of Social Roboticsin Iran, a title she achieved as a Post-Doctoral research associate at the Social Robotics Laboratory of the Sharif University of Technology. Her areas of interest include discourse analysis, interlanguage pragmatics, materials development, and RALL. Dr. Alemi has been the recipient of various teaching and research awards from Sharif University of Technology, Allameh Tabataba'i University, Islamic Azad University, and Int. Conf. on Social Robotics (ICSR-2014).

Ali Meghdari is a Professor of Mechanical Engineering and Robotics at Sharif University of Technology (SUT) in Tehran. Professor Meghdari has performed extensive research in various areas of robotics; social and cognitive robotics, mechatronics, bio-robotics, and modeling of biomechanical systems. He has been the recipient of various scholarships and awards, the latest being: the 2012 Allameh Tabatabaei distinguished professorship award by the National Elites Foundation of Iran (BMN), the 2001 Mechanical Engineering Distinguished Professorship Award from the Ministry of Science, Research & Technology (MSRT) in Iran, and the 1997 ISESCO Award in Technology from Morocco. He is currently the Director of the Centre of Excellence in Design, Robotics, and Automation (CEDRA), an affiliate member of the Iranian Academy of Sciences (IAS), a Fellow of the American Society of Mechanical Engineers (ASME), and the Founder and Chancellor of Islamic Azad University-Fereshtegaan International Branch (for students with special needs; primarily the Deaf and Blind).