Learning Bodily Expression of Emotion for Social Robots through Human Interaction

Nguyen Tan Viet Tuyen, *Student Member, IEEE,* Armagan Elibol, *Member, IEEE,* and Nak Young Chong, *Senior Member, IEEE*

***Abstract*—Human facial and bodily expressions play a crucial role in human-human interaction to convey the communicator's feelings. Being echoed by the influence of human social behavior, recent studies in human-robot interaction (HRI) have investigated how to generate emotional behaviors for social robots. Emotional behaviors can enhance user engagement, allowing the user to interact with robots in a transparent manner. However, they are ambiguous and affected by many factors such as personality traits, cultures, and environments. This paper focuses on developing the robot's emotional bodily expressions adopting the user's affective gestures. We propose the behavior selection and transformation model, enabling the robots to incrementally learn from the user's gestures, to select the user's habitual behaviors, and to transform the selected behaviors into the robot motions. The experimental results under several scenarios showed that the proposed incremental learning model endows a social robot with the capability of entering into a positive, long-lasting HRI. We have also confirmed that the robot can express emotions through the imitated motions of the user. The robot's emotional gestures that reflected the interacting partner's traits were widely accepted within the same cultural group, and perceptible across different cultural groups in different ways.**

***Index Terms*—Human-robot interaction, Affective behaviors, Imitation learning, Cross-cultural evaluation**

I. Introduction

Nonverbal behaviors have an indispensable role in human-human interaction. During the conversation, people communicate through facial and bodily expressions to convey their emotions that may influence social relationships. The connection between human behaviors and emotion has been investigated from the psychological point of view [1], [2], [3]. In social robotics, the social human-robot interaction should be treated similarly as the interaction with another person [4]. Taking into account the role of emotional expressions in human-human interaction, many studies have focused on generating robot emotional behaviors by estimating environmental stimuli and incorporating robot emotional states. Robots' social cues can enhance the social interaction outcomes by allowing humans to interact in a facile and transparent manner [5].

In this paper, we firstly underline the importance of emotional expressions in human-human interaction and in social human-robot interaction. Previous works related to facial and bodily expressions of robots are summarized in I-A. In I-B, we emphasize the importance of the interacting partner's traits on generating expressive robot behaviors. Along the lines, the psychological perspectives about the infant's social development are explained in I-C. In Section II, the proposed behavior selection model is described in detail and is followed by the transformation model in Section III. In Section IV, the two scenarios of interaction and evaluation are conducted to validate the proposed transformation model. This model is further strengthened by the integration of behavior selection and transformation model in Section V through a scenario of long-term HRI. Finally, the experimental results, research contributions, and future work are summarized in the conclusions and future work section.

The authors are with the School of Information Science, Japan Advanced Institute of Science and Technology, Ishikawa, Japan {ngtvtuyen, aelibol, nakyoung}@jaist.ac.jp

A. Related Works

The imitation of human or animal behaviors for robots has received considerable attention [6], [7], which can positively enhance the interaction between a subject (which could be human [8] or animal [9]) and a target robot to some degree. In terms of humanoid robots, the mimicry of human emotional behaviors for producing robots' cues could be broadly classified into facial and bodily expressions.

The MIT Kismet robot [10] perceived a variety of social cues from the environment through visual and audio resources, and it responded to the interacting partner through its eye gaze and facial expressions. In [11], the cultural factors on the robot's affective behaviors were investigated. The dynamic facial expressions derived from East Asian people is transferred to a social robot head for subjective evaluation. The authors confirmed that, compared to the robot associated with standardized universal expressions defined by the theory-driven approach [12], the robot equipped with the skills of cultural expressions outperformed its comparator robot in terms of both recognition accuracy and human-likeness.

In contrast to the robots' facial expressions, bodily expressions have received less attention from HRI researchers [13], even though the potential of affective gestures had been clearly revealed from the psychological literature [3], [14]. Only a few studies were aimed at implementing the theory-driven approach in psychology for robot bodily expressions [15], [16]. By taking into account the contribution of human body movements to the attribution of emotion [2], [3], robots' bodily expressions could be generated, especially for the robots without a dedicated facial articulation. The creation of body movements for the NAO robot in [15] was mainly inspired from Meijer's work [2] and other related psychological studies. The analysis evidenced that the designed bodily expressions displayed on the robot are appropriately conveyed the target

emotions. Similarly, bodily expressions for the Brian 2.0 robot [16] was based on perspectives of social psychology [2], [3] about the connection between human emotion and bodily movements. The experimental results demonstrated that certain human bodily movements representing social emotions could be effectively displayed on the robot. On the other hand, with the data-driven approach, human motion data could be used for generating robots' social cues. In [17], the emotional postures performed by a professional actor and a professional director were recorded by the motion capture system. Then, the expressive key poses were selected through the recognition accuracy of the participants. Afterward, the corresponding robot's emotional poses were carefully positioned to match the original pose of recorded motion data. The experimental results confirmed that bodily postures displayed by the robot could be used to convey emotions during child-robot interaction. Similarly, the UCLIC Affective Posture and Motion Database [18] was used to generate emotional expressions for different robots in [19]. The authors chose the best recognizable ones, and they were mapped into robots with the provided labels through the proposed transformation model. Without using the motion data obtained from the human gesture database, the Tangy robot [20] can observe a demonstrator's social gestures through a one-shot human demonstration. The perceived actions were fed to the proposed imitation system to generate the robot's mimicking gestures. The authors concluded that their proposed approach endowed the Tangy robot with the capability of imitating the interacting partner's social gestures.

B. Importance of Interacting Partner's Traits on Generating Robot Behavior

It is emphasized that social robots should be capable of communicating and interacting with people in a personalized way, adapting and learning social behavior throughout their lifetime [4]. During everyday communications, robots should be able to re-configure their interaction behaviors adapting to environmental stimuli toward increasing the empathy and the engagement of social interaction. Hence, without using stereotyped patterns of behavior produced from a single instance of human motion data [17], [19], [20], robots are required to perceive and learn the interacting partner's behavior through long-term interaction. By sharing the same patterns of behavior with the interacting partner, empathy, defined as "an affective response more appropriate to someone else's situation than to one's own" [21], could be ensured for long-lasting social interaction. On the other hand, it should be noticed that emotional expressions are highly affected by many factors, such as individual personalities and cultures [22]. Thus, robots' behaviors defined by the theory-driven approach [15], [16] may not match the dynamically changing expressions of the interacting partner. Instead, by choosing the appropriate behaviors in alignment with the interacting partner's traits, this strategy ensures that the robot's behaviors conform to the individual traits. There is a strong psychological evidence, known as the "chameleon effect" [23], defined as the tendency to mimic the posture, facial expressions, and verbal and nonverbal behavior of the interacting partners to conform to social norms. The influence of the user's personality traits on the robot's emotional expression was underlined in [8] as the "law of attraction in HRI". The authors examined the effect of KMC-EXPR robot's personalities in social interactions represented by different facial expressions. The results indicated that the partner feels more comfortable when interacting with the robot having a similar personality than those with different personalities. Interestingly, the influences of the robot's behaviors on the interacting subjects were also be confirmed on the interaction between a rat-like robot and a live rat [9]. On the other hand, the capability of dynamically selecting the appropriate behaviors is a strategy for the maintenance of the social relationship throughout day-to-day interaction [24]. The robot's novel behaviors over time can positively contribute to the user's engagement in long-term interaction even though that is not the most appropriate behavior. For this, previous researches outlined in this section provide the empirical evidence for the need of considering the interacting partner's information obtained through long-term HRI to generate the most appropriate social behaviors for robots. Understanding and reflecting the interacting partner's traits to alter the robot's emotional expressions, it is believed that their behaviors could be more acceptable in a variety of social interaction settings [25].

C. Inspiration from Infant Social Development to Generate Robot's Behavior via Long Term HRI

In order to generate the robot's social behavior reflecting the interacting partner's traits through long-term HRI, this research was inspired from the infant social development process, where the infant's interpretation and behaviors are highly affected by their parents through imitative exchanges [26]. In social referencing, the infant is typically the referrer - the individual who seeks and is influenced by the referencing message which is received from the referee - the person doing the influencing. Referees are usually tend to be the infant's parents, specifically their mother [26], [27]. Throughout the infant's social development, they are rapidly influenced by the guidelines from their parents in acquiring knowledge about a typical event. They generate emotions and behaviors in response to the stimuli by an imitating mechanism which regulates their emotions and behaviors to match the encoded emotions and expressions from their parents. Through imitative exchanges, an infant learns a wide variety of skills, customs, and typical behaviors of their culture [28]. All of those play a crucial role in helping infants explore and learn about themselves as well as others as a social being.

The idea of infant social development could be used for long-term HRI, allowing the interacting partner to influence the robot's behaviors throughout the social referencing. During day-to-day interaction, the robot incrementally perceives the individual partner's emotional behaviors as the stimulus, and the robot then utilizes the obtained information to form its own interpretation about the corresponding event. More specifically, our proposed behavior selection model sequentially collects the individual's emotional behaviors corresponding

to the specific emotion. Then, by assessing the frequency of the observed human behaviors, the model outputs the most appropriate patterns of emotional behavior. Finally, through the proposed transformation model, behaviors are converted to the robot's bodily expressions. Fig. 1 illustrates the overall flow of the proposed process. This process is continuously repeated throughout everyday interaction as a social development process of the robot.

II. The Behavior Selection Model

The characteristics and types of human affective behavior vary according to the culture and personality traits of individuals [22]. Therefore, collecting labeled data from human behaviors during social interaction is a challenging task. Unsupervised learning sidesteps the requirements of labeled data to enable robots to be capable of learning socially appropriate gestures based on human behaviors. This idea has been shared across different contexts. In [29], the unsupervised learning approach is presented for the association between human gestural commands and robot actions. In [30], the authors validated the performance of different unsupervised learning algorithms such as Self Organizing Maps (SOM), Fuzzy C-Means (FCM), and K-Means for the recognition of human posture in video sequences. The capability of robot arm trajectory learning from human demonstrations was proposed in [31], where the trajectory clustering and approximation modules take human demonstrative trajectories as the input and then classify these trajectories into different groups. For each group, the most consistent trajectory was selected and a set of generated trajectories can be visualized in a simulated environment, allowing the human user to finally select the desired trajectory. For unstructured HRI with no *a priori* information about human behaviors, unsupervised learning is an effective strategy, allowing robots to acquire new knowledge of the interacting partner's behaviors by classifying various types of actions into different groups based on the similarity of patterns.

On the other hand, through day-to-day social interaction, robots may acquire new knowledge incrementally. This means that robots should be able to learn new information in an incremental manner without corrupting the existing knowledge. This strategy ensures robots to acquire a collection of skills throughout its developmental process. In [32], the authors proposed a system which enables robots to incrementally learn unlabeled gesture patterns based on the interaction with a human partner. In [33], the robot is able to improve its visual perception by incrementally learning from newly detected objects associated with the labels provided by the user through interactions.

The aforementioned studies have shown that unsupervised learning in an incremental manner is a desirable approach for long-term HRI, especially when the number of observed human behaviors continuously increases. The following sections will detail our proposed approach to cope with such situations.

A. Encoding Human Actions into Feature Descriptors

During day-to-day HRI, sequences of human bodily expression data can be obtained for each of the different human emotions. Before feeding this data to the training phase, an appropriate method should be applied to encode the human actions into the feature descriptors. Since human behavior data vary from one behavior to another, the pose estimation module may produce different frame lengths for different type of actions. The encoder should produce the fixed-length descriptors regardless of the number of obtained skeleton frames. The covariance descriptor proposed by [34] can satisfy such a requirement and achieve higher accuracy compared to the other approaches [35].

Let us consider an action $A_i = [S_1, S_2, S_3, ..., S_T]$ performed in the period of time T. The estimated skeleton at frame i $(1 \leq i \leq T)$ is a vector $S_i = [x_1, x_2, ..., x_k, y_1, y_2, ..., y_k, z_1, z_2, ..., z_k]^t$ which represents k joints of the skeleton. Consequently, $N = 3k$ elements are included in the vector S_i. The covariance matrix $C(S)$ is calculated as:

$$C(S) = \frac{1}{t-1}\sum_{i=1}^{t}(S_i - \overline{S})(S_i - \overline{S})^{\intercal} \tag{1}$$

where $\overline{S}$ is the sample mean of S_i computed over the time t and $^{\intercal}$ represents the transpose operator. The upper triangle of $C(S)$ contains $(N \times (N+1)/2)$ elements.

It should be noticed that the covariance matrix described by Eq. 1 only captures the spatial features of the action A_i. By computing the overlapped covariances over the entire time T, the temporal information of motion could be determined. Specifically, a covariance matrix $C(S)$ at the level l is calculated by Eq. 1 that covers $t = T/2^l$ frames. At the top level $(l = 0)$, a single covariance matrix is computed capturing the entire action including T frames. At level $l = 1$, there are 3 smaller overlapping time windows covering $T/2^1$ frames, where the matrices are computed over a period of time $[0, T/2]$, $[T/4, 3T/4]$, and $[T/2, T]$, respectively. Finally, the obtained feature descriptors x_i of action A_i is extracted from the upper triangles of four covariance matrices computed. The vector x_i consisting of $(4\times N\times(N+1)/2)$ elements efficiently represents the spatial and temporal information of the entire sequence A_i. This feature descriptor has been widely used for action recognition in both supervised [34] and unsupervised learning tasks [36].

B. Training and Clustering Phase

Given sets of feature descriptors from the encoding phase, as an unsupervised learning approach without *a priori* knowledge of the number of clusters, Self Organizing Map (SOM) [37] was implemented for the training phase in [38]. SOM creates a set of neurons representing the distributions of the whole dataset, and the topological property of the original data was preserved on the grid of SOM neurons. It should be underlined that topological preservation is the main strength of SOM for classifying the encoded descriptors into different groups based on the similarities. On the other hand, for the scenarios of daily human-robot interaction, since the number of observed behaviors will continuously increase, the robot should be capable of incrementally learning the new gestures

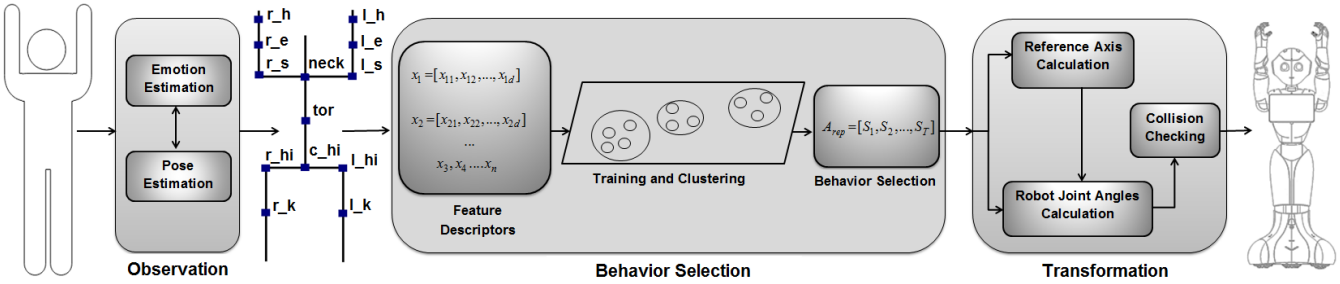


Fig. 1: Social gestures generation framework: The observation part collects information of the partner. The behavior selection part selects the most frequently observed behaviors. The transformation part converts the selected behaviors into robot motions.

without corrupting the existing model. However, with the SOM network, the number of trained neurons must be fixed in advance, which makes this approach inappropriate for incremental learning. To satisfy the requirement of incremental learning for scenarios of day-to-day interactions as well as ensuring the topological preservation, we employ a Dynamic Cell Structure (DCS) neural architecture [36] [39] for the training phase. DCS makes sure that the topological properties will be maintained in a similar way as SOM. Indeed, thanks to the capability of extending the network structure, DCS could learn new patterns in an incremental manner. DCS has been widely used for online learning purposes, such as NASA's first generation Intelligent Flight Control System program [40]. The other type of growing self organizing network named Grow When Required (GWR) [41] was used in [42], where the authors utilized GWR as the supervised learning to recognize the affective states of human bodily expression.

On the DCS network $curI$, for the input descriptor x_i, Eq. (2) is firstly used to determine the closest m_{bmu} and the second closest m_{second} neurons to the descriptor x_i. Then, the lateral connection defining the connection strength between two neurons m_i and m_j is updated by the Hebbian learning rule [43] as described in Eq. (3), where ε is a forgetting constant and ϑ is a threshold for deleting lateral connection.

$$\begin{aligned} ||x_i - m_{bmu}|| &\leq ||x_i - m_i||, \qquad 1 \leq i \leq N \\ ||x_i - m_{second}|| &\leq ||x_i - m_i||, \quad 1 \leq i \neq bmu \leq N \end{aligned} \tag{2}$$

$$C_{ij}(t+1) = \begin{cases} 1 & ,(i = bmu) \wedge (j = second) \\ 0 & ,(i = bmu) \wedge (j \in \{N_i\} \setminus \{second\}) \\ & \quad \wedge (C_{ij} < \vartheta) \\ \varepsilon C_{ij}(t) & ,(i = bmu) \wedge (j \in \{N_i\} \setminus \{second\}) \\ & \quad \wedge (C_{ij} \geq \vartheta) \\ C_{ij}(t) & ,otherwise \end{cases} \tag{3}$$

Similar to SOM, the weight vector of DCS neurons are updated by the Kohonen learning rule [37] which makes them move closer to the current input descriptor x_i as described in Eq. (4), where η is the learning rate, m_i is the neighboring neurons of m_{bmu}, and $\phi(m_i, m_{bmu})$ is the neighborhood kernel function.

$$\begin{aligned} m_{bmu} &= m_{bmu} + \eta(t)(x_i - m_{bmu}) \\ m_i &= m_i + \eta(t)\phi(m_i, m_{bmu})(x_i - m_i) \end{aligned} \tag{4}$$

The resource value τ_{bmu} of the closest neuron m_{bmu} is updated by Eq. (5). The new neuron unit could be added into the network and located between neurons with the largest and second-largest resource values. The training phase is finished by decreasing the resource value τ_i of all neuron units, as described in Eq. 6, where λ is defined as the decreasing rate.

$$\tau_{bmu} = \tau_{bmu} + ||x_i - m_{bmu}||^2 \tag{5}$$

$$\tau_i = \lambda \tau_i \tag{6}$$

It can be seen that when the input data x_i is fed to the training phase, the Kohonen learning rule and the Hebbian learning rule allow the current network $curI$ to modify the lateral connection C_{ij} and the neuron weights m_i. The network is then grown up in an appropriate manner. This process endows the updated network $uptI$ with the capability of preserving the topological property of the whole training data in an incremental way.

After the training phase, m trained neurons are classified into different groups based on its similarities at the training phase. Classifying the trained neurons into different groups is conducted with the distance matrix based approach [44]. Since each trained neuron unit m_i creates a Voronoi region on the original space of feature descriptors x given by Eq. (7), the region V_i may include several descriptors. As a result, by clustering the training neurons rather than the original descriptors directly, it has been reported that significant improvement is obtained in the speed of clustering phase [45]. At the end of the clustering phase, the grid of m neuron units was divided into k clusters. Each neuron m_i and its corresponding descriptor x_i is defined by the Best Matching Unit function given by Eq. (8). Now the descriptor x_i belongs to the same cluster as its corresponding neurons unit m_i.

$$V_i = \{x| \quad ||x - m_i|| \leq ||x - m_j|| \quad \forall j \neq i\} \tag{7}$$

$$||x - m_i|| = min\{||x - m||\} \tag{8}$$

C. Behavior Selection Phase

As explained earlier, n action data $A_1, A_2, .., A_n$ are encoded into n fixed-length descriptors $x_1, x_2, ..., x_n$. Then, during the training and clustering phase, these actions are clustered into k different groups $Cluster_1, Cluster_2, ..., Cluster_k \quad (k \leq n)$ based on the similarities of its motions. At the behavior selection phase, considering the probabilistic distribution of human actions observed by the robot, the most frequently observed behavior is selected out of the largest cluster $Cluster_i (i \in k)$. Here, $Cluster_i$ contains the highest number of patterns sharing the similar features compared to other clusters. As those patterns are repeatedly observed by the framework, they could be seen as the habitual behavior that reflects the interacting partner's traits. Finally, to ensure that the selected pattern geometrically represents the majority of elements in the largest cluster, $Cluster_i$, the representative pattern is defined by Eq. (9). Now the descriptor x_{rep} is the one located closest to the center μ of the $Cluster_i$.

$$||x_{rep} - \mu|| \leq ||x - \mu||, \quad \forall x \in Cluster_i, \tag{9}$$

where $||x - \mu||$ is the Euclidean distance between the center of $Cluster_i$ and the descriptor x. Finally, the corresponding action of descriptor x_{rep} is selected and denoted by A_{rep}. Overall, for a new input action A_i obtained, the behavior selection model is executed as summarized in Algorithm 1. The robot can utilize the interacting partner's habitual action A_{rep} as a reference for generating an appropriate bodily expression associated with a certain emotion.

Algorithm 1 Behavior selection model processing a new observed action A_i.

Input: observed action A_i, current network $curI$, network parameters ϵ, ϑ, η, ϕ, λ;
Output: action A_{rep}, updated network $updI$;

do (action A_i)
  $x_i \leftarrow$ ActionEncoder(A_i);
  $m_{bmu}, m_{second} \leftarrow$ TwoClosestNeurons($curI$, x_i);
  $updI \leftarrow$ HebbianRule($curI$, m_{bmu}, m_{second}, ϵ, ϑ);
  $updI \leftarrow$ KohonenRule($updI$, η, ϕ);
  $updI \leftarrow$ UpdateResource($updI$, m_{bmu});
  $updI \leftarrow$ AddNeuron($updI$);
  $updI \leftarrow$ DecreaseResources($updI$, λ);
  $Cluster_i, \mu \leftarrow$ ClusteringPhase($updI$);
  $x_{rep} \leftarrow$ RepresentativeAction($Cluster_i$, μ);
  $A_{rep} \leftarrow$ ActionDecoder(x_{rep});
end

III. The Transformation Model

Now the pipeline for transferring the selected behaviors to the target robot (SoftBank's Pepper) is explained. It should be emphasized that the number of degrees of freedom (DOFs) and the range of joint angles of the Pepper robot are limited compared to those of humans. Thus, the transformation model is required to convert human actions into Pepper by taking into account its kinematic structure. As shown in Fig. 1, the transformation model receives the human joint vectors $\{r\_k$, l_k, r_hi, l_hi, c_hi, tor, $neck$, r_s, l_s, r_e, l_e, r_h, $l\_h\}$, represented in the Cartesian space. These vectors are represented with respect to the tor coordinates fixed at the torso joint in order to obtain invariant representation to the camera position. Through the proposed model, a set of robot's joint angles are released. For calculating the Pepper robot's joint angles, the solutions to the inverse kinematics problem are computed based on geometric algebra. This approach has been widely used in imitation learning from human behaviors for different kinematic structures of robots such as the DARwIn-OP humanoid robot [46], NAO robot [47], and Tangy robot [20]. Depending on the specific robot's kinematics, an appropriate inverse kinematic model is determined. It can be also noted that there are significant differences in the lower body between humans and the Pepper robot. Therefore, the transformation model focuses on the imitation of the upper body by producing the corresponding movements of the Hip, Head, Shoulders, Elbows, Hands, and Wrists on both the right and left sides.

A. Reference Axis Calculation Phase

The robot exhibits nonverbal communicative behaviors such as head motion in order to convey deeper messages and emotions. Those behaviors affect the orientation of the estimated human pose with respect to the camera embedded on the robot head. To cope with this problem, the reference axes calculated by Eqs. (10), (11), and (12) are used to describe the orientation of the estimated pose. Then, they were combined with the estimated human skeleton to calculate the robot's corresponding joint angles.

$$\overrightarrow{z_{ref}} = (\overrightarrow{r\_hi} - \overrightarrow{l\_hi}) \times (\overrightarrow{r\_hi} - \overrightarrow{tor}) \tag{10}$$

$$\overrightarrow{x_{ref}} = \overrightarrow{r\_hi} - \overrightarrow{l\_hi} \tag{11}$$

$$\overrightarrow{y_{ref}} = (\overrightarrow{z_{ref}} \times \overrightarrow{x_{ref}}) \tag{12}$$

B. Joint Angles Calculation Phase

In order to calculate the robot's joint angle θ corresponding to the movements of the human joints in the Cartesian space, we can define the two vectors, $\overrightarrow{v_1}$ and $\overrightarrow{v_2}$, at each joint, representing the directions of the two neighboring links, respectively. It is straightforward to calculate the joint angle by the dot product between the two neighboring vectors given by Eq. (13). Finally, depending on the robot's home configuration different from that of the human joints, an offset value is added to the calculated θ.

$$\theta = \cos^{-1}\left(\frac{\overrightarrow{v_1} \cdot \overrightarrow{v_2}}{|\overrightarrow{v_1}| \cdot |\overrightarrow{v_2}|}\right) \tag{13}$$

The joint angle calculation phase releases a set of angles Roll (α), Pitch (β), Yaw (γ) corresponding to the availability of DOFs of the robot kinematic structure.

Specifically, a set of joint angles θ_i = { α_{Hip}, β_{Hip}, $\alpha_{RightShoulder}$, $\beta_{RightShoulder}$, $\alpha_{RightElbow}$, $\gamma_{RightElbow}$, $\alpha_{LeftShoulder}$, $\beta_{LeftShoulder}$, $\alpha_{LeftElbow}$, $\gamma_{LeftElbow}$ } are computed. Also, when the human motion capture data of the head and wrists are available, the model additionally generates the robot's joint angles: θ_i = { β_{Head}, γ_{Head}, $\gamma_{RightWrist}$, $\gamma_{LeftWrist}$ }. Details of the calculation are given in the supplementary material.

C. Collision Checking Phase

$$\theta_i = \begin{cases} \theta_{i\_min}, & if \quad \theta_i \leq \theta_{i\_min} \\ \theta_i, & if \quad \theta_{i\_min} < \theta_i < \theta_{i\_max} \\ \theta_{i\_max}, & if \quad \theta_i \geq \theta_{i\_max} \end{cases} \tag{14}$$

The calculated joint angles are checked to ensure that they satisfy the robot's joint limit constraints given by Eq. (14), where θ_{i_min} and θ_{i_max} denote the lower and upper limits of the joint angles θ_i, respectively. Finally, before releasing it to the robot, collision detection is conducted using the Pepper robot's off-the-shelf API to prevent possible self-collisions.

IV. Experiment 1 - Transferring Human Social Gestures into the Robot

In this experiment, the transformation model, which converts human actions into the Pepper robot motions, is qualitatively evaluated in two different scenarios. Firstly, we recruited observers from various cultural backgrounds who are not familiar with robots. They evaluated whether the demonstrators' gestures are appropriately represented by the robot taking into account the robot's physical constraints. Secondly, observers evaluated whether the human emotional expressions were retained by the corresponding robot motions. We performed subjective evaluations widely used to evaluate the robot's facial expressions [10] or bodily expressions [19].

A. Experiment Scenario: Generating Robot Actions through Human Demonstration

1) Experimental Setup: This scenario evaluates the imitated gestures by the robot through a one-shot human demonstration. More specifically, the users stood in front of the Pepper robot to perform 6 different actions. The interacting distance between the demonstrator and the robot was approximately 2 meters. The robot acquired the user's upper body motion as a sequence of skeleton frames using its on-board camera. The pose estimation module receives the human motion as the input, and, through the VNect model [48], a sequence of 3D skeleton frames represented by 14 markers is released. Then, the transformation model sequentially converts demonstrated actions into the robot motion. Additionally, to analyze how similar the actions were performed by the demonstrators, each of the human demonstrated actions H was encoded to the corresponding feature vector C given by Eq. 1. The encoded vector C captures the spatial-temporal information of motions as described in Section II-A. Then, the similarity between a pair of human actions H_a and H_b can be determined by measuring the cosine distance between the two encoded feature vectors C_a and C_b as in Eq. 15. Hence, the closer the cosine distance to 1, the greater the similarity between two vectors.

TABLE I: Similarity between all pairs of human actions.

Action	*H 1*	*H 2*	*H 3*	*H 4*	*H 5*	*H 6*
H 1	1.00	0.63	0.76	0.45	0.07	0.33
H 2	0.63	1.00	0.50	0.47	0.11	0.21
H 3	0.76	0.50	1.00	0.42	0.12	0.39
H 4	0.45	0.47	0.42	1.00	0.20	0.50
H 5	0.07	0.11	0.12	0.20	1.00	0.25
H 6	0.33	0.21	0.39	0.50	0.25	1.00

TABLE II: Confusion matrix representing the recognition of six human actions (H) transferred into the robot model (R), normalized by the number of observers.

Action	Observers					
	H 1	*H 2*	*H 3*	*H 4*	*H 5*	*H 6*
R 1	0.85	0.02	0.13	0.00	0.00	0.00
R 2	0.03	0.94	0.03	0.00	0.00	0.00
R 3	0.13	0.08	0.79	0.00	0.00	0.00
R 4	0.00	0.00	0.00	0.92	0.00	0.08
R 5	0.00	0.00	0.00	0.00	0.92	0.08
R 6	0.00	0.00	0.00	0.05	0.08	0.87

$$Similarity(C_a, C_b) = \frac{C_a \cdot C_b}{||C_a||\,||C_b||} \tag{15}$$

An online survey in English was conducted with 39 observers (28 males and 11 females), ranging in age from 22 to 33 (mean age *M = 25.6* years, standard deviation *SD = 2.5* years), from three different cultures (13 Chinese, 14 Japanese, and 12 Vietnamese). They are graduate students at the Japan Advanced Institute of Science and Technology who use English in daily life. The selected observers are mostly not familiar with robots since their educational backgrounds are not related to robotics and they have not interacted with social robot platforms (such as Nao, Pepper, and others) before. They were asked to evaluate the demonstrator's motions and the Pepper's imitated ones using online surveys discussed further in a later section.

2) Experiment Results and Discussions: The three demonstrators performed six actions combining the movements of their hip and arms, each of them demonstrated two actions. Table I shows the similarity between all pairs of demonstrator's actions calculated from Eq. 1 and Eq. 15. The demonstrators' actions were imitated by the Pepper robot through the transformation model. We conducted a survey with a group of observers using a 23.8-inch color monitor with a resolution of 1920×1080 pixels, in order to evaluate the recognition of demonstrated actions imitated by the robot. The survey form provides a Graphical User Interface (GUI) that help us collect the observers' responses. They were asked to use a keyboard to input their personal information. It is followed by the six experimental trials corresponding to the six different types of the robot actions. On each trial, the observers used a mouse to trigger the video of the Pepper robot's imitated action. After that, they sequentially watched six videos of the human demonstrated actions by triggering one video at one time. The observers used a mouse to select the most similar human action to the robot's one - in a six alternative

forced choice task. Notice that by randomly swapping the positions of the videos, the six human actions were presented to the observers in different temporal orders. This format prohibits the observers from exhibiting a biased response. The duration of each demonstrated action is approximately 6 seconds. The stimuli subtended a visual angle of 11.17° (vertical) and 8.00° (horizontal). The viewing distance is approximately $70\ cm$. Table II shows the recognition rate of the imitated actions, evaluated by 39 observers. It is indicated that the observers could recognize the demonstrators' actions imitated by the robot with the high categorization accuracy. However, the observers were sometimes confused between the human action $H1$ and $H3$. By analyzing the similarity of demonstrators' actions using its encoded feature vectors, Table I confirms that the demonstrated actions $H1$, $H2$, and $H3$ were performed similarly to each other. It should be remarked that the experimental results only show that (1) the robot is able to perceive the user's action represented using a skeleton sequence collected with its on-board sensor and (2) the proposed framework can convert the observed user action into the target robot motion subject to its physical constraints. To evaluate more closely whether the messages of the user's actions are retained by the robot's bodily expressions or not, the transformation model will be validated with the user's affective behaviors detailed in the following experiment.

B. Experiment Scenario: Human Emotional Expressions Retained by Robot Motions

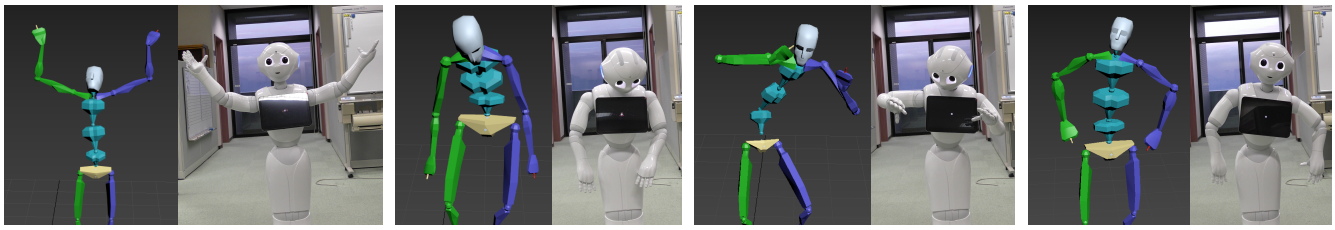
1) Experimental Setup: We conducted a study to evaluate whether the message of human bodily expressions is retained by the robot motions using the UCLIC Affective Posture and Motion Database [18]. The database includes 108 affective gestures recorded by a motion capture system. It is categorized into four emotion labels (*Happy, Sad, Fear, Angry*). The actors conveyed those emotions mostly using their upper body. The acted gestures were evaluated online by 70 subjects from three different cultural groups of observers (25 Japanese, 25 Sri Lankans, and 20 Caucasian Americans in the United States). The evaluation results were represented by the label and the intensity of the emotions. In our experiment, we selected four affective gestures portraying each of the four emotions, respectively, which were recognized correctly by the majority of observers across the above-mentioned cultural groups. Specifically, the selected gestures should satisfy the following two conditions: (1) the sum of percentages of observers across three cultures who correctly recognized the emotion of the gesture is the highest of all the other gestures in the database and (2) on each group, the percentage of observers recognizing the emotion correctly should be equal to or higher than 40%. Here, the threshold of 40% was used to filter out gestures showing a significantly low recognition rate within a specific culture. Finally, the four human gestures were fed to the transformation model to be converted to the robot motions.

Subjective evaluations were carried out through an online survey designed in English. It was conducted with 150 observers (101 male and 49 female), ranging in age from 18 to 45 years old (mean age M = 25.2, standard deviation SD = 4.1 years), from five different cultures (14 Chinese, 11 Japanese, 13 Koreans, 57 Turkish, and 55 Vietnamese). The observers are English speaking students of five universities and institutes, most of whom are not familiar with social robots. Similar to the Experiment IV-A, this survey form is designed with a GUI for collecting the observers' responses. The first part of the survey includes four experimental trials corresponding to the different robot's bodily expressions. The order of trials were randomly presented to the observers. On each trial, the observers were asked to watch the robot's bodily expressions and choose the most appropriate emotion label from the five options ("Happy", "Sad", "Fear", "Angry", "Other") - in a five alternative forced choice task. Here, if the observers believe that the robot's gesture may infer a different message, they select the option "Other" and write their own interpretation. Each of the actions was performed for 7 seconds, and the observers can replay the video as many times as they wish before completing the experimental trial. Another part of this evaluation is the assessment of four selected UCLIC human expressions. The motion capture data were graphically visualized using Autodesk 3ds Max software. Similar to the first part of the survey, there are four experimental trials where their positions are randomly swapped across the observers. The observers were asked to watch the human skeleton actions and rate the emotion label from "Happy", "Sad", "Fear", "Angry", and "Other"- in a five alternative forced choice task. It should be emphasized that, by additionally evaluating the human bodily expressions, this approach allows us to collect the subjective results of human and robot affective gestures which were evaluated by the same group of observers.

2) Experimental Results and Discussions: Figs. 2a, 2c, 2e, and 2g show the key poses of the four selected human expressions (*Happy, Sad, Fear, Angry*) chosen from the UCLIC dataset. Through the transformation model, those gestures were converted to the Pepper robot motions considering the physical constraints as shown in Figs. (2b, 2d, 2f, 2h).

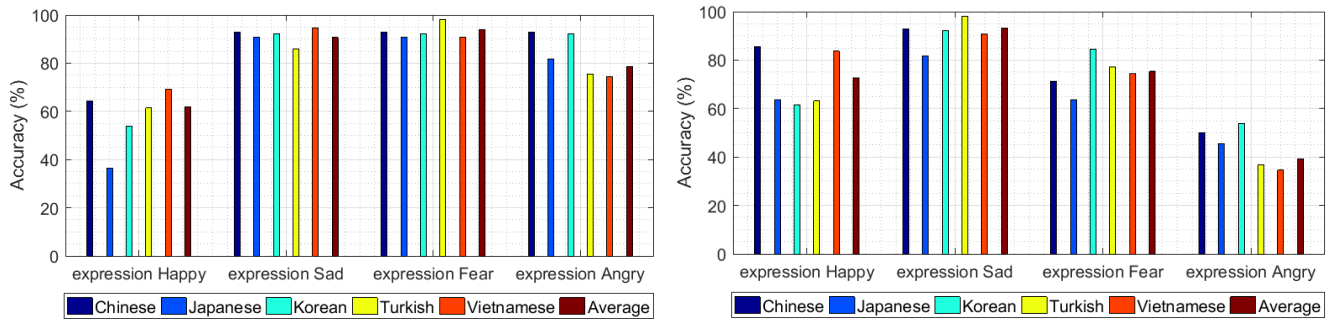
Subjective evaluations were conducted for both the human and robot emotional bodily expressions. Figs. 3a and 3b show the culture-specific recognition accuracy. Additionally, the average recognition accuracy was calculated by pooling data of 150 observers across five cultural groups. It can be seen from Fig. 3a that the overall recognition accuracy of human bodily expressions is quite high. However, only 36% of the Japanese observers correctly recognized the human expression *Happy*. The overall recognition accuracy is also high for the robot bodily expressions (*Happy*, *Sad*, and *Fear*) as seen in Fig. 3b. Notably, the bodily expression *Angry* has the lowest recognition accuracy.

Fig. 2e shows the key pose of the human motion *Fear*. It consists of bending the upper body, covering the face with their hands, and stepping backward to defend themselves. It should be noticed that a coordinated movement of head, shoulder, arms, and knees is required as well as the backward step. Due to the differences in the lower body between the human and the robot, the knee motion and the backward step were removed in the robot motion. As a result, the robot's joint β_{Knee} is set to a constant value of $\beta_{Knee} = 0\ rad$ (as the value at the initial position). Indeed, Fig. 4 indicates the absolute differences



(a) human *Happy* (b) robot *Happy* (c) human *Sad* (d) robot *Sad* (e) human *Fear* (f) robot *Fear* (g) human *Angry* (h) robot *Angry*

Fig. 2: Selected human postures from UCLIC dataset visualized using Autodesk 3ds Max: Figs. 2a, 2c, 2e, and 2g represent the key poses of human bodily expressions. Figs. 2b, 2d, 2f, and 2h show the corresponding Pepper expressions.



(a) Human bodily expressions (b) Robot bodily expressions

Fig. 3: The recognition accuracy of bodily expressions rated by observers within each cultural group. The dark-red bar indicates the average pooled accuracy of 150 observers across five cultures.

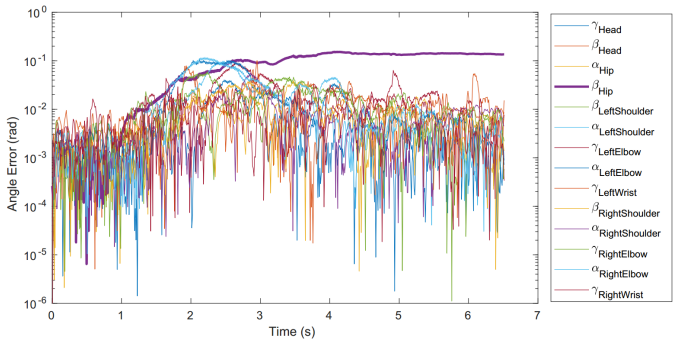


Fig. 4: Absolute differences in joint angle values between the human expression *Fear* and the imitated one performed by the Pepper robot.

between a set of joint angles calculated from the human motion *Fear* and angle values collected from the robot's sensors. It is noticed that the joint β_{Hip} could not reach the desired values of the human motion, due to the limitation of the robot's physical configuration. This error constrains the range of bending motion of the robot's upper body, failing to reach the extent as performed by the human skeleton. These reasons affected the recognition of the robot expression *Fear*. Thus, the robot *Fear* was relatively difficult to recognize with the average recognition accuracy 75% compared to 94% for the human skeleton *Fear*.

As shown in Fig. 2a, the expression *Happy* was performed by raising outstretched arms over the head. Since there are no facial expressions to accompany bodily expressions, this expression of the skeleton model sometimes caused the observers to infer other messages such as *Angry*, *Fear*, or *Shocked*. On the other hand, when this expression was conveyed by the robot, it was more easily recognizable to the observers. After completing the survey, the results were shown to the observers for receiving their feedback. It was self-reported that while watching the robot bodily expressions, they commonly paid more attention to the robot face. By looking at the robot face and bodily expressions at the same time, the observers felt that this behavior might imply *Happy* or *Welcoming*. For that reason, the recognition rate of the robot *Happy* is slightly higher than that of the human skeleton *Happy*. It should be underlined that no eye color was used for the robot emotional expressions. However, the robot face influences the recognition of its bodily expression. Indeed, the facial expression turns out to be significant for the robot expression *Angry*. When transferring this gesture to the robot motion, due to the limitation of its physical constraints, the robot could not move its arms close enough to its hip. This problem led to the difficulty in achieving higher recognition rate of its expression *Angry* as shown in Fig. 3b. On the other hand, the robot face caused the observers to infer positive emotions like *Happy* or other message such as *"Hey, what's up?"*. As a result, 25% of the observers rated other meanings for the robot expression *Angry*. The observers also thought that the robot somehow tried to convey expression Angry by its bodily movement. However, they were confused by the robot face. It should be noted that the design of Pepper's face was influenced by characters in Japanese animation having big eyes [49]. That appearance makes the robot look more friendly to humans even when no animated behaviors are performed by

the robot. Accordingly, the Pepper's face positively contributes to the recognition of *Happy*, while it adversely affected the perception of *Angry*.

Interim Summary: In this experiment, the transformation model was sequentially evaluated by two different experimental setups. In the scenario of learning from human demonstrations, the robot was able to perceive the demonstrators' gestures and imitate them as closely as possible under the robot's physical constraints. The robot's imitated behaviors were recognized with high categorization accuracy. Secondly, the human emotional expressions represented by the robot motions were evaluated using the public dataset. The messages of *Happy*, *Sad*, and *Fear* were well retained by the robot motions. The robot's expression *Angry* was recognized with low accuracy, mainly due to the robot's physical constraints and facial expression.

V. Experiment 2 - Long-Term Interaction to Develop Robot Emotional Expressions

A. The Scenario of Interaction

In this experiment, the transformation model and the behavior selection model were integrated into a new scenario of three consecutive days of HRI. In other words, the proposed framework in Fig. 1 was comprehensively evaluated through a long-term interaction scenario. The experimental setup is given in Fig. 5, where the Pepper robot interacted with a demonstrator to learn from his emotional behaviors. The interacting distance between the user and the Pepper robot was about 2 meters. The interaction section was triggered when the robot detected the user through the facial detection API[1]. Then, the robot started the conversation by executing several verbal and nonverbal behaviors from the predefined list of interacting actions. The demonstrator then responded to the robot with his facial and bodily expressions in his own way since no constraints were placed on them. The human upper body motion is captured from the robot's camera, using the human pose estimation module as described in the previous experiment, and the demonstrator's gestures were acquired as a sequence of 3D skeleton frames represented by 14 markers. At the same time, the robot estimated the user's facial expression through the emotion estimation API[2]. The user's emotional behaviors associated with facial expressions *Happy*, *Sad*, and *Fear* were stored in the robot memory. For each interaction day, the obtained user data were sequentially fed into the corresponding emotion classes in the behavior selection model as presented in Algorithm 1, which was followed by the transformation process. In the next day, the robot gained access to the stored knowledge from the previous day and incrementally learned from the user's new behaviors. The scenario of interaction was repeatedly carried out for three consecutive days, considering the number of interactions obtained and especially the familiarity of the demonstrator with the experimental protocol.

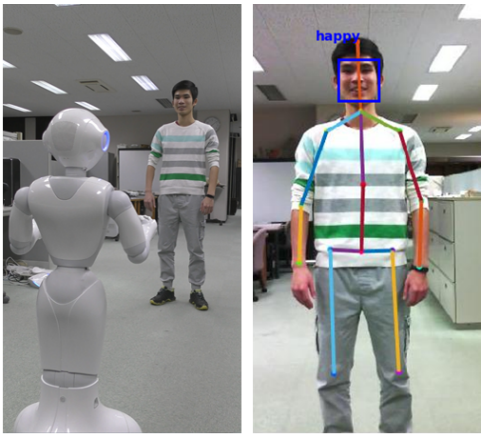


Fig. 5: The scenario of Pepper's interaction for 3 consecutive days learning from the partner's emotional behaviors.

TABLE III: SOM versus DCS on MSRC-12 dataset.

	SOM	DCS
$Precision$	0.9166	0.8019
$Recall$	0.9115	0.9141
F_{value}	0.9133	0.8524

B. Evaluation Criteria

This survey investigates the quality of the robot's emotional behaviors aligned with the interacting user's culture (Vietnamese) as well as the cultural differences in the perception of the robot's behavioral expressions. Specifically, subjective evaluations were performed through an online survey designed in English. We recruited 136 observers (96 males and 40 females), ranging in age from 18 to 45 (mean age M = 25.2 years, standard deviation SD = 4.1 years) from five different cultures (13 Chinese, 9 Japanese, 13 Koreans, 44 Turkish, and 57 Vietnamese). The observers are students from five different universities and institutes. They are fluent in English and most of them are not familiar with robots.

Firstly, the observers were asked to watch the robot's emotional expression and choose the appropriate emotional label similar to the previous experimental setup mention in section IV-B. Then, the observers rated the appropriate value for Arousal and Valence using the Self-Assessment Manikin (SAM) nine-point scale [50]. Arousal and Valence are the dimensions on the Circumplex model of affect [51]. This validation allows the observers to assess and express their emotional responses to the robot's expression without any constraints on the emotion labels. The observers' assessments were then scaled in a range of [-1, 1]. This measurement has been widely used by other HRI researchers to subjectively validate the robot's behaviors [52], [53].

C. Experiment Results and Discussion

1) Robot Bodily Expressions Generated Over Three Consecutive Days of Interaction: The behavior selection model

[1]http://doc.aldebaran.com/2-5/naoqi/peopleperception/alpeopleperception.html

[2]http://microsoft.com/cognitive-services/en-us/

TABLE IV: The behavior selection phase on the third day. Using Eq. (9), the representative pattern A_{rep} is selected as the closest one to the center μ of the largest cluster $Cluster_i$.

(a) $Cluster_i$ on emotion class *Happy*

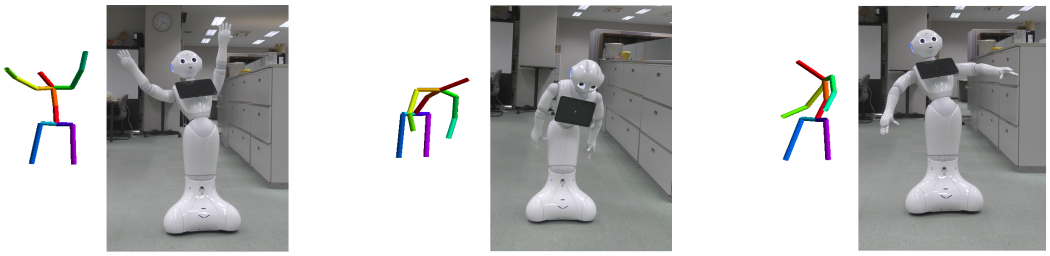
	Pattern ID	$\|\|x-\mu\|\|$
1	H_39	0.3937
2	H_47	0.3042
3	H_45	0.3794
4	H_40	0.3974
5	H_31	0.3370
6	H_5	0.3889
7	H_7	0.3071
8	H_20	0.3152
9	**H_41**	**0.2002**
10	H_23	0.3230
11	H_28	0.2656
12	H_42	0.2992
13	H_50	0.2298
14	H_22	0.3495
15	H_30	0.3342
16	H_13	0.2506
17	H_51	0.2798
18	H_43	0.3533
19	H_32	0.2425
20	H_38	0.2824
21	H_36	0.2440

(b) $Cluster_i$ on emotion class *Sad*

	Pattern ID	$\|\|x-\mu\|\|$
1	S_24	0.1636
2	S_33	0.1828
3	S_14	0.1600
4	S_20	0.1926
5	S_6	0.1917
6	S_4	0.2249
7	S_29	0.3187
8	S_40	0.1326
9	S_39	0.1428
10	S_23	0.1685
11	S_42	0.2099
12	S_17	0.1373
13	**S_27**	**0.1237**
14	S_21	0.1622
15	S_32	0.3039
16	S_18	0.1488
17	S_15	0.1890
18	S_25	0.1900
19	S_38	0.3070
20	S_35	0.4049
21	S_41	0.3948
22	S_30	0.3619
23	S_31	0.3965

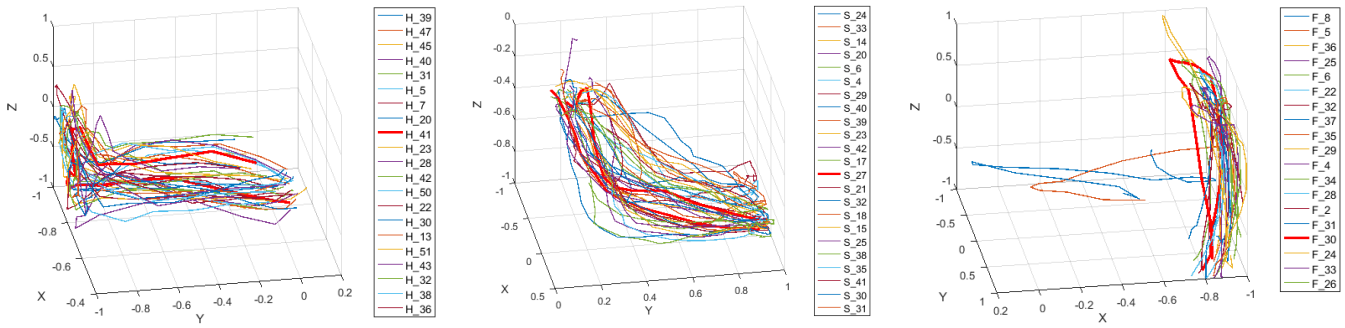
(c) $Cluster_i$ on emotion class *Fear*

	Pattern ID	$\|\|x-\mu\|\|$
1	F_8	0.9811
2	F_5	1.0957
3	F_36	0.5164
4	F_25	0.3147
5	F_6	0.2713
6	F_22	0.3075
7	F_32	0.3386
8	F_37	0.3134
9	F_35	0.2958
10	F_29	0.2819
11	F_4	0.3764
12	F_34	0.3324
13	F_28	0.4791
14	F_2	0.3354
15	F_31	0.2600
16	**F_30**	**0.2451**
17	F_24	0.4249
18	F_33	0.4563
19	F_26	0.3133



(a) Expression *Happy* (b) Expression *Sad* (c) Expression *Fear*

Fig. 6: The key poses of Pepper bodily expression generated using A_{rep} of the behavior selection phase.



(a) Emotion space *Happy* (b) Emotion space *Sad* (c) Emotion space *Fear*

Fig. 7: The trajectories of human left hand created by the patterns of Table IV. Eq. (9) selects the representative gesture A_{rep} the most consistent one in the cluster.

incrementally perceived the interacting user's emotional behaviors. In more detail, on each emotion class, the demonstrator actions were first encoded to feature descriptors. Those descriptors were incrementally trained and clustered into different groups during the training and clustering phase. Through the behavior selection phase, the representative action A_{rep} was selected. Finally, the transformation model converted the selected expressions into the robot motions. This process was continuously repeated over three consecutive days as a part of the robot's social development. Fig. 10 shows the number of learned behaviors and the changes in the robot's emotional expressions over three days. More specifically, Table IV rep-

resents the selected patterns from the behavior selection phase conducted in the last day. Based on the transformation model, the selected behaviors were converted to the robot motions, being the robot's emotional expressions. Fig. 6 shows the key poses of those behaviors.

In our previous work [36], the training and clustering phase as shown in Fig. 1 was evaluated with the Microsoft Research Cambridge-12 Kinect gesture dataset (MSRC-12) [54]. The experiment results as summarized in Table III indicated that SOM yielded better performance than DCS. Notably, the accuracy of DCS was acceptable, whereas the incremental learning gained considerable benefit on the processing time required compared to SOM. Concerning the long-term interaction scenarios, the robot's capability of incrementally updating the learning model without corrupting the existing one is the most demanding requirement as discussed before. Thus, the DCS was finally selected for our training phase.

Through the training and clustering phase, the obtained data were classified into different clusters based on the similarities. At the behavior selection phase, considering the probabilistic distribution of human actions observed by the robot, the largest cluster, $Cluster_i$, was determined. Among the gestures that belong to $Cluster_i$, instead of randomly picking up one pattern out of the cluster, the representative pattern is defined as the gesture closest to the center μ of $Cluster_i$ as described in Eq. (9). Eq. (9) guarantees that the representative gesture A_{rep} is the most consistent one in that cluster. Tables IVa, IVb, IVc show the patterns located in $Cluster_i$ on each of the emotion classes. The selected pattern A_{rep} represents the majority of elements in the largest cluster $Cluster_i$. With the motion patterns defined in Table IV, the trajectories of the human left-hand are depicted in Fig. 7. Here, the movements of the hand were analyzed, since the hand movements are considered as the richest source of emotional body language [55]. Concerning the behavior selection as described in Table IVa, it is easy to notice that pattern H_41 satisfies Eq. (9). Visualizing the trajectories as shown in Fig. 7a, the trajectory created by the gesture H_41 is correctly located in the center of the cluster. As shown in Table IVc, it can be seen that F_30 is the representative pattern, while the calculated distance of F_5 and F_8 are significantly different to the others in this group. The visualization of their trajectories in Fig. 7c explains the differences. Although inappropriate patterns could exist in $Cluster_i$ due to the performance of DCS in the training phase, the behavior selection phase ensures that the selected emotional gesture A_{pre} is the most reasonable one among the others in $Cluster_i$. Those representative actions A_{pre} were converted to the Pepper robot's motions through the transformation model as presented by the key poses in Fig. 6.

2) The Cultural Differences in the Perception of Robot Expressions: While the experiment results in Section IV confirmed the capability of the robot conveying its emotion through bodily expressions, in this experiment, we aim to evaluate the human perception of the robot behaviors across different cultures. The robot gestures on the last day as shown in Fig. 6 were selected for evaluation. It is reasonable to think that those emotional expressions sufficiently reflected the interacting partner's traits. The interacting user agreed that those expressions are his interested behaviors, as he frequently used such gestures to convey his emotion. Thus, the user was easily able to recognize the expressions represented by the Pepper robot. For further investigation on how appropriate the robot's emotional expressions would be from the viewpoint of other people, we recruited observers from five different cultural groups. Table V shows the recognition rate of 57 observers who share the same cultural background with the interacting user (Vietnamese). Then, this group of observers scored the values of Arousal and Valence for the robot behaviors as shown in Fig. 8. Table VI shows the recognition rate of 136 observers across five different cultures, while Figs. 9a and 9b represent the mean of Arousal and Valence assigned by the observers within individual cultural groups.

Table V confirmed the high recognition accuracy of the robot expressions *Happy* rated by Vietnamese observers. 75% of them believed that Pepper tried to convey *Happy* cues by its bodily movements. 11% thought that the gesture means $Fear$. 9% felt that the gesture might have another meaning

TABLE V: The recognition rate of robot expressions rated by 57 observers from the same cultural group with the interacting partner, normalized by the number of observers.

Emotional label	Observers			
	Happy	***Sad***	***Fear***	***Others***
Happy	0.75	0.05	0.11	0.09
Sad	0.05	0.65	0.11	0.19
Fear	0.19	0.02	0.60	0.19

TABLE VI: The recognition rate of robot expressions rated by 136 observers from 5 different cultural groups, normalized by the number of observers.

Emotional label	Observers			
	Happy	***Sad***	***Fear***	***Others***
Happy	0.72	0.03	0.13	0.12
Sad	0.07	0.54	0.13	0.26
Fear	0.13	0.03	0.67	0.17

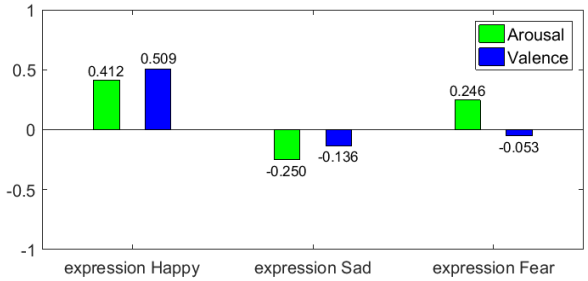
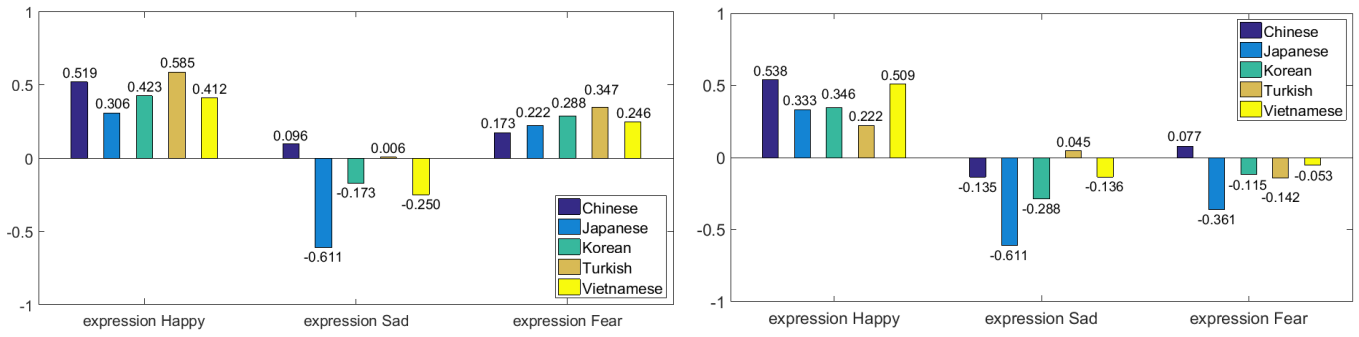


Fig. 8: Mean values of Arousal and Valence rated by Vietnamese observers for robot expressions.

TABLE VII: The differences in Arousal and Valence for expressions *Happy*, *Sad*, *Fear* rated by Vietnamese observers. The third column indicates significantly different pairs.

Dimension	ANOVA test	Post-hoc test
Arousal	**significant diffs.** $\mathbf{p\_value = 1.08E{-}14}$	Sad-Happy = 4.02E-14 Sad-Fear = 5.27E-09
Valence	**Significant diffs.** $\mathbf{p\_value = 3.8E{-}08}$	Happy-Sad = 1.36E-07 Happy-Fear = 9.47E-06



(a) Arousal dimension

(b) Valence dimension

Fig. 9: Mean values of Arousal and Valence rated by people from 5 different cultures.

TABLE VIII: The cultural differences in Arousal and Valence rated by Chinese (CHI), Japanese (JAP), Korean (KOR), Turkish (TUR), Vietnamese (VIE) observers. The third column indicates significantly different pairs.

(a) Arousal dimension

Emotion	ANOVA test	Post-hoc test
Happy	No significant diffs. $p\_value = 0.1610$	No significant diffs.
Sad	**Significant diffs. p_value = 0.0001**	VIE-TUR = 0.0278 TUR-JAP = 0.0012 JAP-CHI = 0.0019
Fear	No significant diffs. $p\_value = 0.7197$	No significant diffs.

(b) Valence dimension

Emotion	ANOVA test	Post-hoc test
Happy	**Significant diffs. p_value = 0.0171**	VIE-TUR = 0.0117
Sad	**Significant diffs. p_value = 0.0028**	VIE-JAP = 0.0431 JAP-TUR = 0.0018
Fear	No significant diffs. $p\_value = 0.2992$	No significant diffs.

such as *Excited*. The Pepper robot expressed the emotion *Sad* by slowly bending its upper body, keeping the hand positions lower than its Hip. 65% of observers assigned the label *Sad* to such Pepper motions. 19% rated it as another label like *Sorry*. The Pepper robot suddenly moved backward and raised its arms forward to express *Fear* which was recognizable to 60% of observers. On the other hand, such energetic movements made 19% of observers confused with *Happy*, or it might cause them to infer another message such as *Shocked*.

Table VI shows the recognition rate of 136 observers from five different cultures. In general, there were no significant differences noticed on the recognition rate of emotion label assigned by Vietnamese observers (who share the same cultural background with the interacting partner) and the others. However, a wide variety of answers about the possible message of the robot's expressions were received from the non-Vietnamese observers. More specifically, the evaluation results indicated that 26% of observers rated expression *Sad* by other labels which have the similar meaning such as *Shy*, *Boring*, or *Uncomfortable*. 12% of observers believed that expression *Happy* might be other positive cues such as *Thankful*, *Cheer*, or energetic expressions like *Excited* or *Euphoric*. These results suggested that the generated expressions were not only recognizable to the observers who have the same cultural background as the interacting partner, but also recognizable to the observers from different cultural groups.

To address in more detail about the differences in the perception of the robot's emotional behaviors, the following discussion focuses on the Arousal and Valence dimensions of the Circumplex model of affect. These dimensions allow us to investigate how the observers perceived the robot's emotional expressions without being affected by the interpretation of emotional labels. Firstly, to analyze the differences within the generated gestures using the Arousal and Valence values assigned by Vietnamese culture as shown in Fig. 8, the one-way analysis of variance (one-way ANOVA) test was conducted in the Arousal dimension. It was followed by analyzing the Valence dimension. When the significant differences were detected from the ANOVA test ($p < 0.05$), the post-hoc test was carried out to explore the differences. Table VII summarizes the obtained results. The ANOVA test indicated that there were significant differences ($F(2, 168) = 39.188, p = 1.08E{-}14 < 0.05$) in the Arousal dimension of the three generated behaviors. Then, the post-hoc test revealed that the Arousal values for *Sad* was significantly different with *Happy* ($p = 4.02E{-}14 < 0.005$) and *Fear* ($p = 5.27E{-}09 < 0.05$). Thus, the observers from this cultural group assigned similarly the Arousal values for *Happy* and *Fear* higher than *Sad*. Likewise, the significant differences were also found in the Valence dimension ($F(2, 168) = 18.947, p = 3.8E{-}08 < 0.05$). Analyzing the post-hoc test, these significant differences come from *Happy-Sad* ($p = 1.36E{-}07 < 0.005$) and *Happy-Fear* ($p = 9.47E{-}06 < 0.005$). Consequently, the results revealed that the observers rated similarly higher values of Valence for *Happy* than *Sad* and *Fear*. It is widely known that Arousal represents the energy of emotion, while Valence describes the extent to which an emotion is positive or negative. Hence, it can be inferred that the observers from this culture tended to perceive the robot expression *Happy* with a positive emotion than the robot expression *Sad* and *Fear*. In contrast, they thought that Pepper performed *Happy* and *Fear* more energetically than *Sad*.

Figs. 9a and 9b represent the mean values of Arousal and Valence, respectively, rated by 136 observers across five different cultures. To analyze how different the Arousal and Valence values are within these cultures on each emotion class, the ANOVA test was conducted with the Arousal and Valence dimensions. Once the significant differences were detected ($p < 0.05$), further analysis with the post-hoc test was carried out to determine which pair of cultures are significantly different from each other. Tables VIIIa and VIIIb summarize the analysis on the Arousal and Valence dimensions, respectively. Firstly, the results indicated that Vietnamese observers were more likely to rate lower Arousal than those who were Turkish for the robot expression *Sad*. Also, Japanese observers tended to assign lower values of Arousal than the Chinese and Turkish observers for *Sad*. In the Valence dimension, Vietnamese observers rated higher values than Turkish for *Happy*. On the other hand, the Japanese observers were more likely to assign lower values for *Sad* than those who were Vietnamese and Turkish. Hence, the differences in perception of robot emotional behaviors have been clearly distinguished on the Arousal and Valence dimensions. More precisely, the Vietnamese observers tended to feel *Happy* more positively than the Turkish observers. In contrast, those who were Vietnamese felt that the robot expression *Sad* was performed less energetically than the way those who were Turkish perceived. Similarly, Japanese observers seemingly thought that *Sad* was expressed less intensively than the Turkish and Chinese cultural groups. At the same time, Japanese observers were more likely to think that Pepper conveyed more negative emotion than the way Vietnamese and Turkish observers perceived. In general, the significant differences as mentioned above suggested that different cultural groups perceived the same emotional expressions of the robot in different ways.

Interim Summary: In this experiment, the scenario of long-term HRI was conducted to validate the proposed incremental learning approach. In the behavior selection model, the training and clustering phase was revisited. Then, the role of the behavior selection phase for selecting the representative patterns was emphasized. Through the transformation model, the patterns were converted into the robot motion. Subjective evaluations were conducted to evaluate how appropriately the emotional expressions were represented by the robot. A series of validations were conducted in the emotion label categories as well as on the Arousal and Valence dimensions. The evaluation results indicated that the robot behaviors, which reflected the interacting partner's traits, are easily recognizable to the group of observers who share the same cultural background with the partner. The results also support the notion that the robot gestures are recognizable and perceptible to the observers of other cultural groups in different ways.

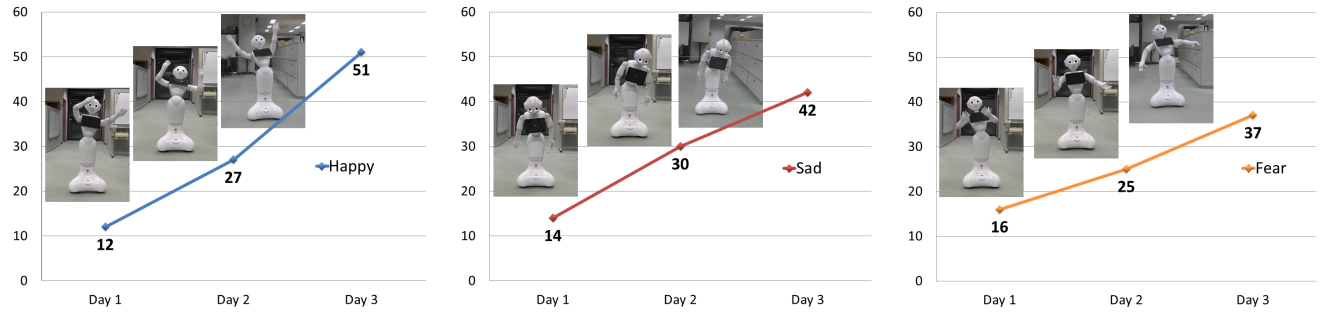
VI. Conclusions and Future Work

This work presented the human behavior selection and transformation framework to autonomously generate emotional bodily expressions for social robots. Compared to the existing studies, the proposed framework enables the robot to be capable of obtaining the individual partner's habitual behaviors through long-term human-robot interaction, leading to generating the robot's social gestures. The overall approach was inspired by the social development of infants, where the behaviors and interpretation of infants are highly influenced by their parents. Similarly, our approach emphasizes the role of the interacting partner's traits to generate the robot's social behaviors. A series of experiments were designed to verify the effectiveness of the proposed behavior learning strategy. Firstly, the human behavior transformation model allowed the robot to learn from the interacting partner's one-shot demonstration. Then, the model was validated using a publicly available human affective posture and motion dataset. The experimental results revealed that the robot was able to generate imitated human behaviors. Furthermore, the message of human emotional expressions was well retained by the robot's behaviors. Secondly, the behavior selection model and the transformation model were integrated into a scenario of long-term social interaction. Through the interaction over three consecutive days, the robot produced the emotional bodily expressions which reflected the interacting partner's behaviors. These expressions were evaluated by observers from different cultural groups. The experimental results confirmed that the robot's emotional expressions were widely recognizable to the people sharing the same cultural background with the interacting partner. Likewise, the robot expressions were recognizable and perceptible to different cultural groups in many different ways. The current results also support the psychological findings that emotional behaviors are affected by many different factors such as individual personalities and cultural backgrounds. Therefore, by acquiring and reflecting the interacting partner's behaviors through long-term interaction, social robots are endowed with the capability of incrementally learning to develop their social behaviors to adapt to its environmental settings.

Our contributions in this paper are: (1) the incremental learning approach to the representative behavior selection reflecting the interacting partner's traits over long-term interaction, (2) the transformation model to convert human behaviors into the target robot's motion space, and (3) the dataset of human expressions obtained from the robot's point of view that could be used for other researches in the field of emotional gesture recognition. Summarizing, the proposed approach enables social robots to develop and learn their behaviors to adapt to a variety of social and environmental settings. In our future work, the current nonverbal interaction behavior will be extended to support more sophisticated behavior of social robots associated with the verbal content of emotional speech.

Acknowledgment

This work was supported by the EU-Japan coordinated R&D project on "Culture Aware Robots and Environmental Sensor Systems for Elderly Support," commissioned by the Ministry of Internal Affairs and Communications of Japan and EC Horizon 2020 Research and Innovation Programme under grant agreement No. 737858. The authors are also grateful for financial support from the Air Force Office of Scientific Research under AFOSR-AOARD/FA2386-19-1-4015.



(a) Emotion space *Happy* (b) Emotion space *Sad* (c) Emotion space *Fear*

Fig. 10: Variational patterns of emotional behavior obtained through 3 consecutive days: Pepper robot incrementally learns and updates its bodily expressions day by day.

REFERENCES

[1] C. Darwin and P. Prodger, *The expression of the emotions in man and animals*. Oxford University Press, 1998.

[2] M. De Meijer, "The contribution of general features of body movement to the attribution of emotions," *Journal of Nonverbal Behavior*, vol. 13, no. 4, pp. 247–268, 1989.

[3] H. G. Wallbott, "Bodily expression of emotion," *European Journal of Social Psychology*, vol. 28, no. 6, pp. 879–896, 1998.

[4] C. L. Breazeal, *Designing sociable robots*. MIT Press, 2004.

[5] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics and Autonomous Systems*, vol. 42, no. 3-4, pp. 143–166, 2003.

[6] A. J. Ijspeert, "Biorobotics: Using robots to emulate and investigate agile locomotion," *Science*, vol. 346, no. 6206, pp. 196–203, 2014.

[7] Z. Gao, Q. Shi, T. Fukuda, C. Li, and Q. Huang, "An overview of biomimetic robots with animal behaviors," *Neurocomputing*, vol. 332, pp. 339–350, 2019.

[8] E. Park, D. Jin, and A. P. del Pobil, "The law of attraction in human-robot interaction," *International Journal of Advanced Robotic Systems*, vol. 9, no. 2, p. 35, 2012.

[9] Q. Shi, H. Ishii, S. Kinoshita, A. Takanishi, S. Okabayashi, N. Iida, H. Kimura, and S. Shibata, "Modulation of rat behaviour by using a rat-like robot," *Bioinspiration & Biomimetics*, vol. 8, no. 4, p. 046002, 2013.

[10] C. Breazeal, "Emotion and sociable humanoid robots," *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 119–155, 2003.

[11] C. Chen, L. B. Hensel, Y. Duan, R. A. Ince, O. G. Garrod, J. Beskow, R. E. Jack, and P. G. Schyns, "Equipping social robots with culturally-sensitive facial expressions of emotion using data-driven methods," in *IEEE International Conference on Automatic Face & Gesture Recognition*, 2019, pp. 1–8.

[12] P. Ekman, E. R. Sorenson, and W. V. Friesen, "Pan-cultural elements in facial displays of emotion," *Science*, vol. 164, no. 3875, pp. 86–88, 1969.

[13] R. Hortensius, F. Hekele, and E. S. Crossy, "The perception of emotion in artificial agents," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 4, pp. 852–864, 2018.

[14] H. Aviezer, Y. Trope, and A. Todorov, "Body cues, not facial expressions, discriminate between intense positive and negative emotions," *Science*, vol. 338, no. 6111, pp. 1225–1229, 2012.

[15] M. Häring, N. Bee, and E. André, "Creation and evaluation of emotion expression with body movement, sound and eye color for humanoid robots," in *IEEE International Symposium on Robot and Human Interactive Communication*, 2011, pp. 204–209.

[16] D. McColl and G. Nejat, "Recognizing emotional body language displayed by a human-like social robot," *International Journal of Social Robotics*, vol. 6, no. 2, pp. 261–280, 2014.

[17] A. Beck, L. Cañamero, A. Hiolle, L. Damiano, P. Cosi, F. Tesser, and G. Sommavilla, "Interpretation of emotional body language displayed by a humanoid robot: A case study with children," *International Journal of Social Robotics*, vol. 5, no. 3, pp. 325–334, 2013.

[18] A. Kleinsmith, P. R. De Silva, and N. Bianchi-Berthouze, "Recognizing emotion from postures: Cross-cultural differences in user modeling," in *International Conference on User Modeling*, 2005, pp. 50–59.

[19] G. Van de Perre, M. Van Damme, D. Lefeber, and B. Vanderborght, "Development of a generic method to generate upper-body emotional expressions for different social robots," *Advanced Robotics*, vol. 29, no. 9, pp. 597–609, 2015.

[20] T. Zhang, W.-Y. Louie, G. Nejat, and B. Benhabib, "Robot imitation learning of social gestures with self-collision avoidance using a 3d sensor," *Sensors*, vol. 18, no. 7, p. 2355, 2018.

[21] M. L. Hoffman, *Empathy and moral development: Implications for caring and justice*. Cambridge University Press, 2001.

[22] A. Kleinsmith, P. R. De Silva, and N. Bianchi-Berthouze, "Cross-cultural differences in recognizing affect from body posture," *Interacting with Computers*, vol. 18, no. 6, pp. 1371–1389, 2006.

[23] T. L. Chartrand and J. A. Bargh, "The chameleon effect: the perception–behavior link and social interaction." *Journal of Personality and Social Psychology*, vol. 76, no. 6, p. 893, 1999.

[24] I. Leite, C. Martinho, and A. Paiva, "Social robots for long-term interaction: a survey," *International Journal of Social Robotics*, vol. 5, no. 2, pp. 291–308, 2013.

[25] B. Bruno, N. Y. Chong, H. Kamide, S. Kanoria, J. Lee, Y. Lim, A. K. Pandey, C. Papadopoulos, I. Papadopoulos, F. Pecora *et al.*, "Paving the way for culturally competent robots: A position paper," in *IEEE International Symposium on Robot and Human Interactive Communication*, 2017, pp. 553–560.

[26] S. Feinman and M. Lewis, "Social referencing at ten months: A second-order effect on infants' responses to strangers," *Child Development*, pp. 878–887, 1983.

[27] S. Feinman, D. Roberts, K.-F. Hsieh, D. Sawyer, and D. Swanson, "A critical review of social referencing in infancy," in *Social referencing and the social construction of reality in infancy*. Springer, 1992, pp. 15–54.

[28] A. Meltzoff, "The role of imitation in understanding persons and developing theory of mind," *Understanding Other Minds: Perspectives from Autism*, pp. 335–366, 1993.

[29] Y. Mohammad, T. Nishida, and S. Okada, "Unsupervised simultaneous learning of gestures, actions and their associations for human-robot interaction," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009, pp. 2537–2544.

[30] K. K. Htike and O. O. Khalifa, "Comparison of supervised and unsupervised learning classifiers for human posture recognition," in *International Conference on Computer and Communication Engineering*, 2010, pp. 1–6.

[31] J. Aleotti and S. Caselli, "Robust trajectory learning and approximation for robot programming by demonstration," *Robotics and Autonomous Systems*, vol. 54, no. 5, pp. 409–413, 2006.

[32] S. Okada, Y. Kobayashi, S. Ishibashi, and T. Nishida, "Incremental learning of gestures for human–robot interaction," *AI & Society*, vol. 25, no. 2, pp. 155–168, 2010.

[33] S. Valipour, C. Perez, and M. Jagersand, "Incremental learning for robot perception through hri," *arXiv preprint arXiv:1701.04693*, 2017.

[34] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations," in *International Joint Conference on Artificial Intelligence*, vol. 13, 2013, pp. 2466–2472.

[35] J. R. Padilla-López, A. A. Chaaraoui, and F. Flórez-Revuelta, "A discussion on the validation tests employed to compare human action recognition methods using the msr action3d dataset," *arXiv preprint arXiv:1407.7390*, 2014.

[36] N. T. V. Tuyen, S. Jeong, and N. Y. Chong, "Incremental learning of human emotional behavior for social robot emotional body expression," in *International Conference on Ubiquitous Robots*, 2018, pp. 377–382.

[37] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[38] N. T. V. Tuyen, S. Jeong, and N. Y. Chong, "Learning human behavior for emotional body expression in socially assistive robotics," in *International Conference on Ubiquitous Robots and Ambient Intelligence*, 2017, pp. 45–50.

[39] J. Bruske and G. Sommer, "Dynamic cell structure learns perfectly topology preserving map," *Neural Computation*, vol. 7, no. 4, pp. 845–865, 1995.

[40] M. Perhinschi, G. Campa, M. Napolitano, M. Lando, L. Massotti, and M. Fravolini, "A simulation tool for on-line real time parameter identification," in *AIAA Modeling and Simulation Technologies Conference and Exhibit*, 2002, p. 4685.

[41] S. Marsland, J. Shapiro, and U. Nehmzow, "A self-organising network that grows when required," *Neural Networks*, vol. 15, no. 8-9, pp. 1041–1058, 2002.

[42] N. Elfaramawy, P. Barros, G. I. Parisi, and S. Wermter, "Emotion recognition from body expressions with a neural network architecture," in *ACM International Conference on Human Agent Interaction*, 2017, pp. 143–149.

[43] T. Martinetz, "Competitive hebbian learning rule forms perfectly topology preserving maps," in *International Conference on Artificial Neural Networks*, 1993, pp. 427–434.

[44] J. Vesanto and M. Sulkava, "Distance matrix based clustering of the self-organizing map," in *International Conference on Artificial Neural Networks*, 2002, pp. 951–956.

[45] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," *IEEE Transactions on Neural Networks*, vol. 11, no. 3, pp. 586–600, 2000.

[46] J.-H. Lee *et al.*, "Full-body imitation of human motions with kinect and heterogeneous kinematic structure of humanoid robot," in *IEEE/SICE International Symposium on System Integration*, 2012, pp. 93–98.

[47] I. Rodriguez, A. Astigarraga, E. Jauregi, T. Ruiz, and E. Lazkano, "Humanizing nao robot teleoperation using ros," in *IEEE-RAS International Conference on Humanoid Robots*, 2014, pp. 179–186.

[48] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt, "Vnect: Real-time 3d human pose estimation with a single rgb camera," *ACM Transactions on Graphics*, vol. 36, no. 4, p. 44, 2017.

[49] A. Pandey and R. Gelin, "A mass-produced sociable humanoid robot: pepper: the first machine of its kind," *IEEE Robotics & Automation Magazine*, vol. 25, no. 3, pp. 40–48, 2018.

[50] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

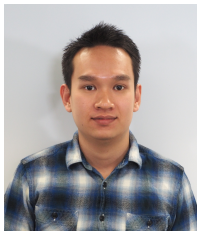
[51] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development and Psychopathology*, vol. 17, no. 3, pp. 715–734, 2005.

[52] M. Marmpena, A. Lim, and T. S. Dahl, "How does the robot feel? perception of valence and arousal in emotional body language," *Paladyn, Journal of Behavioral Robotics*, vol. 9, no. 1, pp. 168–182, 2018.

[53] C. Tsiourti, A. Weiss, K. Wac, and M. Vincze, "Designing emotionally expressive robots: A comparative study on the perception of communication modalities," in *ACM International Conference on Human Agent Interaction*, 2017, pp. 213–222.

[54] S. Fothergill, H. Mentis, P. Kohli, and S. Nowozin, "Instructing people for training gestural interactive systems," in *ACM SIGCHI Conference on Human Factors in Computing Systems*, 2012, pp. 1737–1746.

[55] F. Noroozi, C. A. Corneanu, D. Kaminska, T. Sapinski, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *CoRR*, vol. abs/1801.07481, 2018. [Online]. Available: http://arxiv.org/abs/1801.07481



Nguyen Tan Viet Tuyen received the B.Sc degree in Mechatronics from Ho Chi Minh University of Technology and Education (HCMUTE), Vietnam in 2015, and the M.Sc degree in Information Science from Japan Advanced Institute of Science and Technology (JAIST) in 2018. Currently, he is a Doctoral student at Robotics Lab in School of Information Science, JAIST. His research interests include social and cognitive robotics, human-robot interaction, machine learning, and mechatronics.

Armagan Elibol is a full-time faculty member (as an assistant professor) of Robotics Lab. at School of Information Science of Japan Advanced Institute of Science Technology (JAIST) since October 2018. Prior to that, he was an assistant professor at Department of Mathematical Engineering of Yildiz Technical University (YTU) and was an adjunct teaching faculty member of Computer Engineering Department of FMV Isik University in Turkey. He obtained a Ph.D. in Computer Science (2011) from the Computer Vision and Robotics Group of University of Girona, Spain, both B.Sc. and M.Sc. in Mathematical Engineering from YTU, respectively 2002 and 2004. He was a DAAD-supported visiting researcher at the Pattern Recognition and Bioinformatics Group of Martin-Luther-University Halle-Wittenberg in Germany during summer 2012. Afterward, He was with the Ocean Robotics and Intelligence Lab (ORIN). of Korea Advanced Institute of Science and Technology (KAIST) in Daejeon for one year as a postdoctoral researcher. During 2015, he was with the School of Integrated Technology of Yonsei University at Yonsei International Campus in Songdo-Incheon, Rep. of Korea. His research interests include the wide area of Computer Vision and Robotics. Lately, he integrates his experience on optical mapping to the humanoid research using different sensors and intelligent methodologies.

Nak Young Chong received the B.S., M.S., and Ph.D. degrees from Hanyang University, Seoul, Korea, in 1987, 1989, and 1994, respectively. From 1994 to 2003, he was with Daewoo Heavy Industries, KIST, MEL, and AIST. In 2003, he joined the faculty of JAIST, where he currently is a Professor of Information Science. He also served as a Councilor and Director of the Center for Intelligent Robotics. He was a Visiting Scholar at Northwestern University, Georgia Tech, University of Genoa, and CMU, and served as an Associate Faculty at UNLV and Kyung Hee University. He serves as Editor of the IEEE Robotics and Automation Letters, International Journal of Advanced Robotic Systems, and Journal of Intelligent Service Robotics, and served as Editor of IEEE ICRA, IEEE CASE, IEEE Ro-Man, and UR, and Associate Editor of the IEEE Transactions on Robotics. He served as Program (Co)-Chair for JCK Robotics 2009, ICAM 2010, IEEE Ro-Man 2011/2013, IEEE CASE 2012, URAI 2013/2014, DARS 2014, ICCAS 2016, IEEE ARM 2019. He was a General Co-Chair of URAI 2017 and serves as a General Chair of UR 2020. He also served as Co-Chair for IEEE RAS Networked Robots TC from 2004 to 2006, and Fujitsu Scientific System WG from 2004 to 2008.