Incremental Learning from Humans through Long-term Interaction toward Generating Non-verbal Behavior of Social Robots

Nguyen Tan Viet Tuyen

Japan Advanced Institute of Science and Technology

Doctoral Dissertation

Incremental Learning from Humans through Long-term Interaction toward Generating Non-verbal Behavior of Social Robots

Nguyen Tan Viet Tuyen

Supervisor : Professor Nak Young Chong

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
Information Science

March, 2021

Abstract

People use a wide range of non-verbal channels, including facial and bodily expressions, to communicate their emotions or intention during human-human interaction. Those modalities encourage the communicators' messages could be transmitted to interacting partners in a facile and transparent manner. Being echoed by the influence of human social behaviors, recent studies in human-robot interaction have investigated how to generate non-verbal behaviors for social robots in a way that is appealing and familiar to human partners.

However, non-verbal behaviors are ambiguous. The way how humans express and interpret social behaviors is highly affected by many different factors, including individual personality, cultural background, and other environmental settings. To tackle this problem, the study presented in this dissertation focuses on developing robots' social gestures to adapt to interacting partner's behaviors, allowing generated robots' gestures are familiar to the current social norm. The proposed approach concentrates on the body channel for expressing robots' emotional states and supporting semantic contents of robots' speech. To achieve that, we design the model for generating emotional gestures, the model for generating communicative gestures, and the transformation model. The suggested frameworks endow a robot with capabilities of learning from human behaviors obtained through long-term interaction and transforming generated gestures into the robot's motion, being the robot's social cues supporting for different interaction contexts. We demonstrated the proposed idea on a target social robot. A series of experiments was conducted to evaluate the designed frameworks considering the human perception of generated robot's social cues and the quality of generated gestures. The experimental results also confirmed that different users may interpret the same robot's gesture in different ways. Therefore, the problem of behavior adaption should be addressed when designing non-verbal cues for social robots.
Keywords: social robots, human-robot interaction, non-verbal behaviors, emotional gestures, communicative gestures, imitation learning.

Acknowledgments

This dissertation briefly summarizes a fascinating journey which is a truly life-changing experience for me. It would be impossible for me to accomplish this arduous trip without the huge support and guidance that I received from many people.

I would like to take this opportunity to express my sincere gratitude to my advisor, Professor Nak Young Chong for proving me with an opportunity to start this journey. I am grateful that I was assigned to the HRI research group and joining the CARESSES project, where I had a great chance to work with other outstanding researchers to contribute to the development of social robots. There were many unpredictable challenges throughout the last 3 years of my Ph.D., thank you very much for providing generous support, insightful advice, and kind encouragement to overcome all the obstacles towards pursuing the degree.

I would like to send my special thanks to Professor Armagan Elibol for providing me valuable feedback that pushed me to sharpen my thinking and brought my work to a higher level. I am indebted to Professor Sungmoon Jeong who proving me with insightful guidance at the beginning time when I started to involve in the social robotics domain. I also wish to express my sincere gratitude to Professor Antonio Sgorbissa, Professor Kazuhiro Ogata for giving constructive comments and suggestions on my work during the off-campus minor research. Indeed, it would be impossible for me to complete this dissertation without the assistance of Professor Masato Suzuki, Professor Nguyen Le Minh, Professor Shogo Okada, Professor Ho Seok Ahn and again, Professor Antonio Sgorbissa. The comments and advice from the committee members helped me greatly to improve the clarity of my dissertation.

Throughout 5 years living in Japan, I always received kind encouragements and supports from JAIST friends that push me to keep moving forward. In particular, I would like to express my special thanks to all members of Professor Chong's lab and the Vietnamese community for supporting me in both research and daily life. I am also very grateful to all the JAIST staffs for supporting me throughout

the time studying at JAIST. I will always remember about 5 years living in this beautiful country with the best memories.

Most importantly, I would like to express my deepest thanks for the endless love and unwavering support from my family at all times. To my parents and my brothers, thank you very much for encouraging me in all of my pursuits and inspiring me to follow my dreams. Family is always the peaceful anchor for me during the storm days. I wish to dedicate this work to them.

Committee Members

Member and Supervisor	Professor Nak Young Chong Japan Advanced Institute of Science and Technology, Japan
Member	Professor Nguyen Le Minh Japan Advanced Institute of Science and Technology, Japan
Member	Associate Professor Shogo Okada Japan Advanced Institute of Science and Technology, Japan
Member	Associate Professor Antonio Sgorbissa University of Genoa, Italy
Member	Associate Professor Ho Seok Ahn University of Auckland, New Zealand

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In human-human interaction, people use a wide range of non-verbal behaviors to communicate their emotions or intentions to interacting partners. Among those modalities, facial and bodily expressions play a crucial role. They encourage messages of the communicators could be transmitted to partners in a facile and transparent manner. The connections between human non-verbal behaviors and emotions or speech have been explored in early works [4, 5, 6, 7]. For conveying emotional states through non-verbal channels, most of the previous works investigated the contribution of human facial features to the appearance of emotion. For instance, Facial Action Coding System (FACS) [4] is a well-known approach for modeling human facial expressions. In term of emotional bodily expressions, the association between body movements and emotions are investigated in [5, 6]. Lastly, the use of body channel for supporting verbal communication has been highlighted in [7]. Understanding the crucial role of non-verbal behaviors in social interaction, a growing interest has been seen in developing social robots' non-verbal behaviors in a way that is appealing and familiar to human interacting partners.

1.1 Research Problem

Social robotics is a subfield of robotics focusing on communicating with people through social interactions. It is common to consider that the social human-robot interaction should be treated in a human-like way, where the interaction with

robots is like the interaction with another person [8]. Being encoded by human social behaviors, considerable attention has been paid to generate non-verbal cues for social robots towards enhancing empathy and user engagement of social interaction. Recently, social robots such as Pepper and NAO are equipped with capabilities of performing human-like gestures supporting daily interaction. However, such robots' gestures are manually designed in advance by animation experts to ensure familiarity and human-likeness. On the other hand, by implementing theories of human standardized facial or bodily expressions, robots' non-verbal behaviors [9, 10] can also be created in a human-like shape. Finally, throughout single-shot demonstration or short-term interaction, it is straight forward to transform human non-verbal behaviors into robots' motion space [11, 12, 13], being robots' social cues. However, it is important to emphasize that human non-verbal behaviors are ambiguous and affected by user personality, cultural background, and other environmental settings [14, 15]. Those factors highly influence how people interpret the messages encoded in others' facial or bodily expressions. Similarly, in social robotics, the effects of human traits or cultures on human perception of robots' behaviors have been investigated [16, 17, 18]. It is suggested that by employing non-verbal cues defined from theories of standardized human behaviors or created by animation experts, messages encoded in robots' behaviors may not be recognizable to the interacting social norm. Meaning that interacting partners may misinterpret if they are unfamiliar with such non-verbal cues. Likewise, using the one-shot demonstration approach, robots' gestures may not match the dynamically changing behaviors of interacting partners. Indeed, such stereotyped behaviors could not positively contribute to the user's engagement in long-term interaction [19]. To tackle this problem, it is suggested that robots should be capable of gathering the interacting partner's information obtained through long-term interaction to develop their non-verbal skills. By understanding and sharing similar behaviors with the interacting partner, empathy, defined as "an affective response more appropriate to someone else's situation than one's own" could be ensured in social human-robot interaction.

1.2 Overview of Research Approach

In order to solve that problem, the proposed approach would endow robots capability of learning from human behaviors obtained through long-term interaction in an unsupervised manner. In the end, robots are able to produce their own social cues reflected information obtained from the interacting partners. Overall, our approach is inspired by the social development of infants, where the behaviors of infants are highly influenced by their parents. Similarly, the proposed approach emphasizes the role of human behaviors towards generating robots' non-verbal cues. This approach guarantees the influence of interacting partners on robots' gestures. Vice versa, generated robots' behaviors would be familiar to the interacting social norm.

The proposed approach focuses on generating bodily expressions for social humanoid robots. However, it is noticed that bodily movements could be used to signal a variety of messages during interactions. This research emphasizes the use of bodily expressions for two common purposes: (1) conveying robots' emotional states (in this dissertation, such gestures are called emotional gestures), and (2) supporting concrete contents (known as *iconic* in human behavior studies) or abstract meaning (known as *metaphoric*) of robots' speech (here, they are called communicative gestures).

Since emotional gestures are connected to internal states while communicative gestures are correlated to contents of speech. It is required to design two different models of gesture generation, allowing each of them could be treated in the most appropriate way. Indeed, collecting labeled data from human behaviors during social interaction is a challenging task. Thus, the two designed frameworks are equipped with the capability of learning a sequence of human behaviors in an unsupervised manner. At the generation phase, appropriate gestures are outputted to express certain contexts. Through the designed transformation model, those actions are transformed into the target Pepper robot, taking into account robot physical constraints, and being the robot's social gestures. The generated non-verbal cues can be used in different scenarios of social human-robot interaction.

1.3 Contribution to the Non-verbal Behavior Generation of Social Robots

With the growth of interests in social robotics, it is becoming difficult to ignore the role of non-verbal behaviors when designing social companion robots. There are several interesting approaches and promising ideas for generating robots' social behaviors supporting interactive communication. However, taking into account the problem of behavior adaption that has not been addressed efficiently in previous works, our proposed approach could positively contribute to this domain. Specifically, comparing to off-the-shelf modules embedded into social robots such as Pepper [20], Nao [21] or several interesting models of behavior generation [22], [23], [24], [25], [26]. Rather than programming a set of robots' gestures in advance (by animation experts to ensure the human-likeness of robots' actions) and establishing a set of rules for parameterizing contexts of interaction. Our proposed frameworks for generating gestures endow robots capability of learning human social behaviors, collected from interactions, in an unsupervised manner. This approach does not require prior knowledge of experts to handcraft robot gestures and parameterize models of behavior generation as previous works. Indeed, our solution allows generated robots' gestures are familiar with interacting partners. On the other hand, comparing to the other interesting ideas allowing humans to teach robots new gestures through demonstration [27], [12], [28], [29], [30], [31]. In our proposed approach, rather than using a single shot of demonstration to finalize the robot's emotional or communicative gestures, robots are provided capability of incrementally learning from human behaviors and dynamically adapting their behaviors throughout long-term interaction. The proposed approach focuses on the generation of robots' bodily expressions supporting social interaction. As the result, it could be used for a wide range of social robots, especially the ones without dedicated facial articulation such as NAO, Pepper, Romeo, RoboThespian, and so on.

1.4 Dissertation Outline

In this dissertation, we focus on generating non-verbal behaviors for social robots through imitating human behaviors. In chapter 1, we highlight recent research ideas in this domain and draw several concerns on developing non-verbal cues in social robotics as the research motivation. Then, we explain an overview of the proposed approach to address this research problem.

In chapter 2, we present an overview of human non-verbal behaviors, focusing on the use of face and body channels to signal human intention during communication. It is followed by discussing the important roles of non-verbal cues in social robotics, and the influence of social interaction settings on human perception of robots' behaviors. Finally, the infant social development process is briefly described as an inspiration to develop robots' social skills.

In chapter 3, we illustrate the framework to generate emotional gestures.

In chapter 4, we describe the framework to generate communicative gestures.

In chapter 5, we address the problem of transforming generated human-like gestures into the target social robot.

In chapter 6, a series of experiments is conducted to validate the proposed frameworks in chapter 3, 4, and 5 as well as integrations among them.

In chapter 7, we summarize the research results, contributions, and future directions to improve and extend the current work.

Chapter 2

Research Background and Motivation

In this chapter, we firstly provide an overview of human non-verbal behaviors. In particular, we focus on the use of gestures to convey human emotion and support the semantic contents of human speech during communication. It is followed by emphasizing the role of non-verbal cues in social robotics. However, social behaviors are ambiguous and affected by many different factors. Rather than implementing theories of human standardized behaviors for social robots, the need of considering interacting partners' traits and other factors when generating robots' social gestures will be addressed in 2.2.2. Finally, psychological perspectives about infants' social development are explained in 2.2.3. This idea could be applied to social robots, providing them capable of interacting with human partners and develop their social behaviors.

2.1 Human Non-verbal Behaviors

2.1.1 Emotional Gestures

Emotional expression is one of the most important characteristics in human-human interaction [32]. It has been shown that human emotion and physical expressions are highly associated with each other. During social interaction, people communicate through facial and bodily expressions [33], messages encoded in their affective

Table 2.1: Body movements and postures accompanying specific emotions

Emotion	***Description***
Happiness	Jumping, dancing for joy, clapping of hands, during excessive laughter whole body is throw backwards and shakes
Sadness	Motionless, passive, head hangs on contracted chest
Pride	Head and body held erect
Shame	Turning away the whole body
Fear	Head sinks between shoulders, motionless or crouches down
Anger	Whole body trembles, intend to push or strike violently away

behaviors are used to convey their emotions that may influence social relationships. In the well-known work of Darwin [5], the authors investigate how specific emotions could be interpreted via different behavioral modalities such as facial and bodily features. Table 2.1 presents general movement protocols accompanying specific emotion of Darwin's work [5], which was summarized by Wallbott in his article [6]. In term of facial expression, the author [34] draw the first attention on how human facial muscles change the visual appearance of their face. Based on those findings, the Facial Action Coding System (FACS) [4] is a well-known approach for modeling human facial expressions, including several basic emotions (happiness, disgust, fear, surprise, anger, and sadness). In this coding system, facial expressions can be broken down into individual components of the movement, call Action Unit (AU). As the result, a collection of certain AUs provides information about which emotion is being expressed. Table. 2.2 shows combinations of AUs to form basic emotions.

It is noticed that emotions could be expressed by a wide range of non-verbal channels such as eye movements, facial expressions, and bodily expressions [35]. However, the majority of research on emotional non-verbal has focused on facial expressions while bodily expressions have been lagged so far behind [36]. According to De Gelder [37], about 95 percent of literature in this domain focus on facial expressions as a source for emotion analysis. Most of the remaining 5 percent have been carried out with other modalities while bodily expressions comprising the smallest number of studies. A question has been raised about the role of body expression as a reliable non-verbal channel to convey the emotional states of communicators. In [38], the experimental results demonstrated that emotions could be

Table 2.2: Facial expressions and associated Action Units.

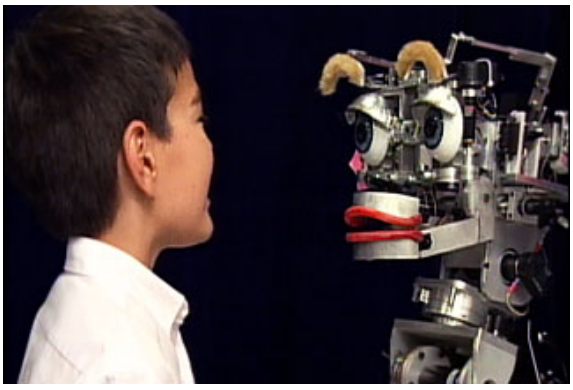
Emotion	***Description***
Happiness	Cheek Raiser (AU6), Lip Corner Puller (AU12)
Disgust	Nose Wrinkle (AU9), Lip Corner Depressor (AU15) Lower Lip Depressor (AU16)
Fear	Inner Brow Raiser (AU1), Outer Brow Raiser (AU2) Upper Lid Raiser (AU4), ...
Surprise	Inner Brow Raiser (AU1), Outer Brow Raiser (AU2) Upper Lid Raiser (AU5), ...
Anger	Upper Lid Raiser (AU5), Lid Tightener (AU7) Lip Tightener (AU23), ...
Sadness	Inner Brow Raiser (AU1), Brow Lowerer (AU4) Lip Corner Depressor (AU15)

determined in videos of body gestures without speech or facial expressions. On the other hand, when observers are presented with affective displays containing a combination of facial expressions and posture or body movement, the authors [36][39] concluded that bodily movements may provide more information than facial expressions for distinguishing between fear and anger or fear and happiness. Indeed, Mehrabian [40] found that the communicator's attitude toward interacting partners highly affected by body configuration and orientation. In conclusion, the above-mentioned studies convince that in addition to facial expression, bodily expressions can be implemented as an important modality for emotional expressions as well as interpretation of affect.

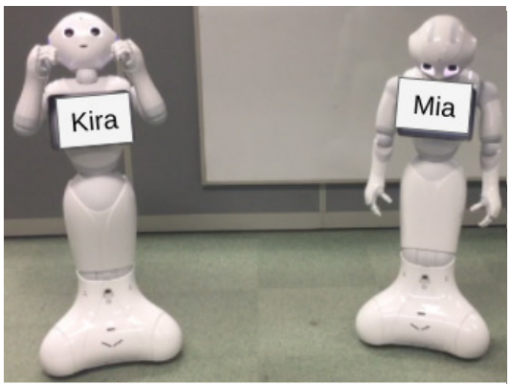
2.1.2 Communicative Gestures

During social interaction, people use non-verbal channels not only for conveying their emotional states, but also supporting for verbal communication. In particular, gestures allow messages, encoded in communicators' speech, could be transmitted to interacting partners (listeners) in a facile and transparent manner [41]. At the same time, interacting partners are well-attentive to information conveyed through such non-verbal behaviors [42]. According to McNeill [7], co-speech gestures and speech are simultaneously generated from a common thought source. When a communicator produces a message to his or her interacting partners, most

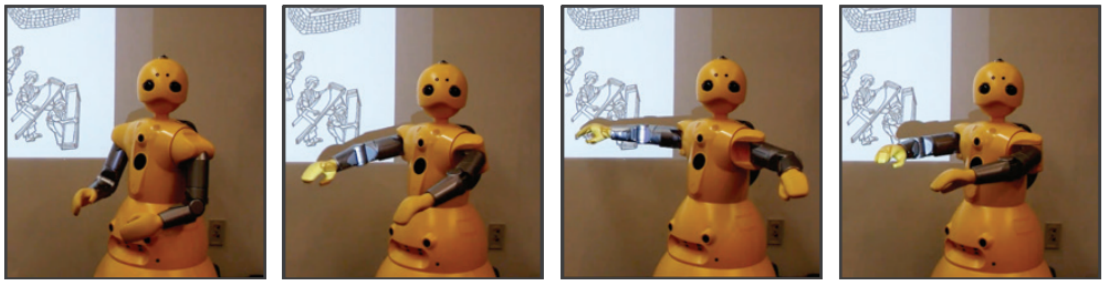
of the information that they want to share is described in their speech. The rest of the information is encoded in their gestures. McNeill [7] categorized communicative gestures into four types: iconic, metaphoric, deictic, and beats as shown in Table 2.3. Iconic gestures have a close connection with semantic contents of the communicator's speech. For instance, when someone says "The ball is very big", they may spread out their two hands to convey how big the ball is. Similarly, when saying "I was driving the car when you called me", he or she may do a "steering wheel" gesture while saying "driving car". Iconic gestures are used to describe movements/shapes of objects/people in space [43], they are concretely connected to semantic contents of communicators' speech. On the other hand, metaphoric gestures are very similar to iconic gestures. However, they are utilized to express abstract concepts rather than concrete meaning (as in the case of iconic gestures). Deictic gestures are also known as pointing gestures. This type of gesture is used to refer to something by pointing with hands or fingers. For example, when someone says "your phone is one that desk", they may point toward the phone or the desk. In this context, the deictic gesture is used to concretely express the speaker's speech. However, this type of gesture can be applied for pointing abstract concepts. Lastly, beat gestures are defined as rhythmic movements of hands. Rather than conveying semantic information of speech, beat gestures are used to stress specific keywords or phrases of speech. This type of gesture is connected to speech prosody. According to McNeill [7], beat gestures are frequently used by human speakers. For instance, in a video corpus of people narrating the events from a Tweety cartoon, the frequency of using beat gestures is 44.7% [7] . Finally, it is important to notice that the meaning of co-speech gestures is freely designed by communicators. Co-speech gestures are capable of expressing a full range of meaning arisen from communicators.



(a) Kismet's emotional facial expression.



(b) Pepper's emotional bodily expression.



(c) Wakamaru's iconic gesture to convey *"a wooden or stone surface"*.

Figure 2.1: Non-verbal cues displayed on Kismet, Pepper, and Wakamaru robot.

Table 2.3: Four types of co-speech gestures and their functions.

Gesture type	***Description***
Iconic	Connect to semantic contents of speech.
Metaphoric	Connect to abstract concept rather than concrete context.
Deictic	Refer to something (e.g., around communicators) by pointing with hand, finger, etc.
Beat	Rhythmic movements for stressing important words No semantic connection to context of speech

2.2 Robot Non-verbal Behaviors

2.2.1 Role of Non-verbal Cues in Social Robotics

Social robotics is a subfield of robotics focusing on communicating with people through social interactions. Here, the social human-robot interaction should be treated in a human-like way, where the interaction with robots is like the interaction with another person [8]. According to the study [44], social robots should be equipped capability of transmitting signals to interacting partners to provide feedback of their internal states and allow humans to interact with them in a facile and transparent manner. For social robots, interactive modalities could be facial expressions [45, 46], bodily expressions [47, 48, 49], eye gaze [49], proxemics [50], and so on. In the following part, we highlight the role of social robots' facial and bodily expressions in previous studies.

The MIT Kismet robot [45] is a well-known work, which made the robot enter into neutral and intuitive social interaction for learning and interacting. The Kismet robot is able to observe a variety of stimuli from the surrounding environment through visual or audio channels. Then, this robot signals their feedback to interacting partners through eye-gaze and facial expressions. Noticed that this robot can display a variety of facial movements by blending several basic prototypical facial expressions along *Arousal*, *Valance*, and *Stance* axes in affect space. On the other hand, in [47], the authors investigated the role of culture in representing robot emotions. Similar to the theory of infant social development [51] where information is injected by humans during the early stage of development and subject to change through long-term interaction. The study showed that robots can learn to behave socially in alignment with individual cultural identity. In their experiment, bodily expressions were utilized to convey robots' emotional states generated by the proposed emotion mechanism. As displayed in Fig. 2.1b, the results conveyed that under different effects of culture, different robots (named Kira and Mia, respectively) could generate different emotional responses towards the same environmental stimuli. The messages encoded in the robots' bodily expressions well present those differences. In the study [48], the authors designed a narration scenario to model associations between human narrators' gestures and

semantic contents of their speech. It is followed by manually implemented the collected narrators' gestures into the target Wakamaru robot. Their experimental setup enabled the target robot capability of performing deictic, beat, iconic, and metaphoric gestures in a narration scenario to a group of participants. Taking into account subjective evaluation results, the authors confirmed that robots' co-speech gestures positively affected participants' information recall, and ability to retell the robot story. Based on that finding, the authors suggested that robots' co-speech gestures could be applied for educational purposes to improve student learning. A similar result has been reported in [52], where iconic gestures performed by a tutoring agent have been shown to improve learners' memory performance.

In short, the aforementioned studies have convinced that it is difficult to ignore the role of non-verbal cues in designing social robots. By utilizing non-verbal modalities, especially facial and bodily expressions, for conveying robots' emotional state as well as supporting robots' speech, it is suggested that robots could interact with users in a facile and transparent manner.

2.2.2 Role of Interacting Partner's Behaviors on Generating Robots' Social Behaviors

It is important to emphasize that social robots should be capable of communicating and interacting with people in a personalized way, adapting and learning social behavior throughout their lifetime [8]. During everyday communications, robots should be able to re-configure their interaction behaviors adapting to environmental stimuli toward increasing empathy and engagement of social interaction. By sharing the same patterns of behavior with interacting partners, empathy, defined as "an affective response more appropriate to someone else's situation than to one's own" [53], could be guaranteed for social human-robot interaction. On the other hand, it is noticed that human verbal and non-verbal behaviors are highly affected by many factors, such as personality, which is a set of distinctive characteristics among humans or cultures, which is shared characteristics of a group of people [14, 15]. Thus, by using theories of human standardized bodily or facial expressions mention in section 2.1 to generate robots' non-verbal behaviors, it is suggested that robots' behaviors may not be recognizable or familiar to the current

environmental setting. In [16], cultural factors on the perception of robots' facial expressions have been investigated. The experiment was conducted on a robot equipped with "universal" and the one with "culturally-derived" facial expressions. By using the theory of human standardized facial expressions [4], the Fuhat robot is equipped with skills of universal facial expressions. On the other hand, facial features derived from East Asian people are transferred to the target robot, being the robot's culturally-sensitive expressions. Subjective evaluation conducted with East Asian participants showed that the robot equipped with skills of cultural expressions (derived from East Asian) outperformed its comparator robot in terms of both recognition accuracy and human-likeness.

In human psychological studies, there is evidence, known as the "chameleon effect" [54]. It is defined as the tendency to mimic the posture, facial expressions, and verbal and nonverbal behavior of the interacting partners to conform to social norms. A similar strategy should be applied for social robots, allowing them to re-configure their interaction behaviors adapting to environmental stimuli toward increasing empathy and engagement of social interaction. In the human-robot interaction domain, it is also known as the "law of attraction in HRI" [17]. This finding was demonstrated by an interesting experiment in [17]. The authors examined the influence of KMC-EXPR robot personality which was reflected by facial expression (features of extrovert and introvert were displayed on the robot through facial movement, size of the robot face/mouth, and eye contact). The experiment results showed that, in terms of friendliness and social presence, extroverted participants considered the extroverted robot more friendly and more socially present than the introverted robot. Vice versa, the introverted participants preferred the introverted robot. Based on that finding, the author indicated that partners fee more comfortable when interacting with the robot having a similar personality than those with different personalities. A similar finding was confirmed in Aly's work [18]. It was shown that extroverts prefer the robot that performing dynamic gestures than introverts. Based on the experimental results, the authors suggested that the problem of human-robot personality matching should be addressed in creating robots' behaviors. Finally, it is noticed that the capability of dynamically selecting the appropriate behaviors is a strategy for the maintenance of the social relationship throughout day-to-day interaction [19]. The robot's novel be-

haviors over time can positively contribute to the user's engagement in long-term interaction.

Previous studies outlined in this section provide the empirical evidence for the need of considering the interacting partner's information obtained through long-term human-robot interaction to generate the most appropriate social behaviors for robots. Understanding and reflecting the interacting partner's traits to alter the robot's gestures, it is believed that their behaviors could be more acceptable in a variety of social interaction settings [55].

2.2.3 Inspiration from Infant Social Development Process for Developing Robots' Social Behaviors

According to psychological researches of human behaviors, one of the most common things that humans do is that gathering their desired information from the surrounding environment and then utilizing it to form their own interpretation and behaviors. Once, the individual has become interested in some environmental events, they are always receptive to information about this event and pay attention to it as soon as it is provided [56]. In the article [57], the author showed that human behaviors are often influenced by social referencing, meaning that humans tend to use the perception and interpretation of another person's to form their own knowledge about specific events. That is the typical way how infants acquire new social skills for their social development. In social referencing, an infant typically is a referer - the individual who seeks and influenced by referencing messages which are received from referees - the person doing the influencing, the referees are always the infant's parents, especially mother [51]. An infant is rapidly influenced by the guideline from their parents in acquiring knowledge about typical events. They generate emotion and behavior in response to the stimuli by an imitating mechanism that regulates their own emotions and behaviors to match the encoded emotions and expressions from their parents. An interesting example was mentioned in [57] where the 9-month-old infant sees that his father plays with a novel toy. The infant infers that his father likes the toy because he smiles. Then, the infant may assimilate this favorable interpretation which can influence her/his behavior when given an opportunity to play with the toy in the future. The capa-

bility to learn through imitation becomes a powerful and flexible form for infants in their social development. Through imitative exchanges, an infant can learn a wide variety of skills, customs and typical behavior of their culture [58] which plays a crucial role in helping the infant explore and learn about themselves and others as a social being. An infant can imitate a wide variety of acts in various scenarios such as facial expressions, gestures, object-related actions, etc. The infant social development process is an interesting idea that could be implemented for social robots, allowing them to incrementally develop their non-verbal behaviors through social interactions with specific human partners. A robot (play a role as an infant) observes interacting partners' social behaviors as their desired stimuli, through imitative exchanges, the robot learns from the acquired information to form their own non-verbal behaviors. Using this approach, the influence of interacting partners on robots' behaviors is guaranteed. Vice versa, generated robots' behaviors would be familiar to the interacting social norms.

Chapter 3

Generation of Emotional Gestures using Dynamic Cell Structure

In this chapter, the model for generating emotional gestures is illustrated. It is started by an overview of related works, focusing on the use of body movements for emotional expression. Then, we highlight several aspects that were not considered in previous studies and describe the proposed approach to tackle those issues. Finally, the designed framework is described in detail. Experiments conducted to validate this framework can be found in section 6.2 of chapter 6.

3.1 Related Works

Facial ad bodily expressions are the two most important modalities to convey the communicator's emotion during human-human interaction. Being echoed by the influence of human social behavior, in social robotics, many studies have focused on generating robot emotional behaviors by estimating environmental stimuli and incorporating robot internal states. Concerning studies about robots' facial expression, the MIT Kismet [45] is a well-known robot which is able to perceive a variety of environmental stimulus and then react to interacting partners through eye gaze and facial expressions. By using the interpolation approach on the three-dimensional affect space - *Arousal*-textitValance-textitStance, the Kismet robot can generate various facial expressions by blending several basic prototype facial

postures together.

In contrast to the robots' facial expressions, studies about bodily expressions have received less attention from the human-robot interaction community [59], even though the potential of affective gestures had been clearly revealed in human behavioral studies [6, 60]. Several studies [9, 10] are motivated by the theory-driven approach. Specifically, by taking into account the contribution of human body movements to the attribution of emotion [61, 6], robots' bodily expressions could be generated, especially for the robots without a dedicated facial articulation. In [9], emotional gestures for the NAO robot are motivated by Meijer's work [61] and other psychological findings [62, 63] on the human expression and perception of emotions. In their experiment, based on the subjective evaluation conducted on *Pleasure-Arousal-Dominance* affect space [64], the authors confirmed that their designed bodily expressions for the NAO robot are recognizable to the subjects. The authors suggested that, so far, theories of human standardized or "universal" bodily expression are applicable for humanoid robots. A similar approach to generate robots' affective gestures can be found in [10]. The implementation of affective gestures for the Brian 2.0 robot is inspired by the previous works of Wallbott [6], and Meijer [61]. In the experiment, the subjective evaluation was conducted to validate the feasibility of their designed robot's emotional bodily expressions from human perception. Similar to [9], the experimental results suggested that certain messages encoded in human emotional gestures are retained effectively on a life-size human-like robot.

It should be emphasized that the way people express and interpret social behaviors is highly affected by many factors, such as cultures, individual personality as discussed in chapter 2. As the result, robots' emotional behaviors [9, 10] implemented from theories of human 'universal' affective behaviors may not match the social norms that robots involve in. This problem could be solved with the data-driven approach, utilizing motion data of particular users to generate robots' social gestures. In [65], by using the emotional postures performed by a professional actor and a professional director, the authors [11] selected the six expressive key poses and then matched them into the NAO humanoid robot, being the robot's emotional poses. The experimental results confirmed that bodily postures displayed by the robot could be used to convey emotions during child-robot interaction. Indeed,

it is shown that the positions of the robot's head highly affected the way humans recognize the robot's emotional states. Similarly, the UCLIC Affective Posture and Motion Database [66] was used to produce robot bodily expressions in [12]. UCLIC dataset includes a set of human affective gestures recorded by a motion capture system. In [12], several patterns of UCLIC dataset were selected out based on the recognition rate. Through the proposed transformation model, those human affective gestures were mapped into the target robots. The subjective evaluation through an online survey showed that generated gestures for certain robot configurations well resemble human gestures. Rather than using motion data obtained from human affective behavior datasets, the Tangy robot implemented in [13] can observe human gestures through a one-shot human demonstration. The obtained data was injected into the proposed framework to generate the robot's imitated gesture. The authors confirmed that the proposed framework endowed the target robot with the capability of observing the interacting partner's social gestures and produce the imitated action taking into account the robot's joint configuration.

Although the information about interacting partners has been taken into account in the aforementioned studies [11, 12, 13]. However, it should be noticed that a single interaction for imitation may not capture the complex of human affective behaviors which can only be observed through long-term interaction. Instead, social robots should be capable of communicating and interacting with people in a personalized way, adapting and learning social behavior throughout their lifetime [8]. To tackle this problem, rather than using a single instance of human motion data [11, 12] or one-shot interaction [13] to finalize the robot affective gestures, our proposed approach provides social robots capability of perceiving and learning interacting partners' behaviors through long-term interaction. The following section will discuss our proposed approach in detail.

3.2 Research Approach

The proposed approach for generating robot emotional body expressions was inspired by the infants' social development. In order to increase the engagement of the conversation and the empathy between a robot and a human through social interactions, robot emotional expressions should conform to the social norm. In

order words, those behaviors should be familiar with users in the current environmental setting. According to human behavioral studies, one of the most common things that humans do is gathering information from the surrounding environment and then utilizing it to form their own interpretation and behaviors [56]. That is the way how infants acquire the new interpretations for their social development [57]. The infant social development process is an interesting idea that could be applied to social robots. This approach allows interacting partners to influence and reconfigure robots' behaviors through long-term social interaction. During day-to-day interaction, the robot incrementally perceives the individual partner's emotional behaviors as their desired stimuli, and the robot then utilizes the obtained information to form its own interpretation of the corresponding event. More specifically, the designed framework 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, human behaviors are converted to the robot's bodily expressions, being the robot's emotional gestures. Fig. 3.1 illustrates the overall flow of the proposed process. This process is continuously repeated throughout everyday interaction as a social development process of the robot.

3.3 Framework Architecture

The characteristics and types of human affective behavior vary according to the culture and personality traits of individuals [14]. 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 [67], the unsupervised learning approach is presented for the association between human gestural commands and robot actions. In [68], 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

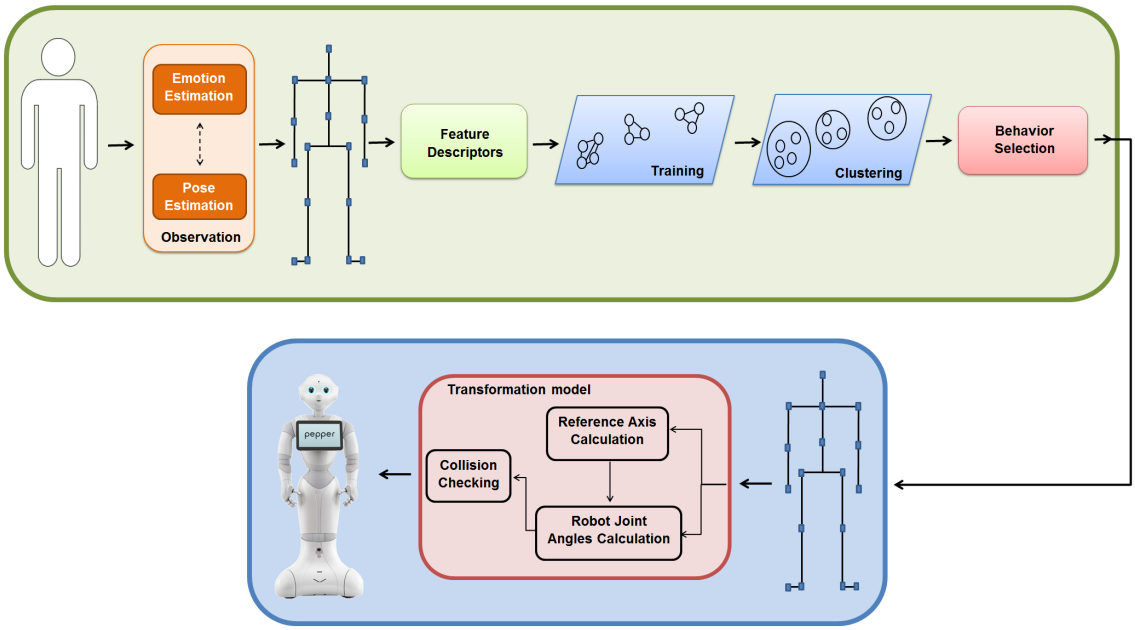


Figure 3.1: The proposed framework for generating emotional gestures: The observation part collects information about the interacting partner. The behavior selection part selects the most frequently observed behaviors. The transformation part converts the selected behaviors into robot motions.

proposed in [69], 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. In summary, for unstructured scenarios of human-robot interaction with no *a priori* information about human behaviors, unsupervised learning is an effective strategy. It allows 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. It means that robots should be able to learn new information incrementally without corrupting the existing knowledge. This strategy ensures robots to acquire a collection of skills throughout its developmental process. In [70], the authors proposed a system that enables robots to incrementally

learn unlabeled gesture patterns based on the interaction with a human partner. In [71], 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.

In short, the previous studies mentioned above have shown that unsupervised learning in an incremental manner is a desirable approach for long-term interaction, especially when the number of observed human behaviors continuously increases. The following sections will detail our designed framework to cope with such situations.

3.3.1 Feature Descriptor

During social human-robot interaction, the observation module collects human bodily expression data and associating them with the estimated emotion. Consider that $A_i = [S_1, S_2, S_3, ..., S_T]$ is the human action collected from the robot's pose estimation module. A_i is a sequence of skeleton frame Si $(1 \leq i \leq T)$ performed in a period of time T. Each frame S_i captures k joints of human motion in 3D space. Before feeding the obtained bodily expression A_i into the training phase, an appropriate method should be implemented to encode the raw data A_i into a motion feature vector x_i. It is straight forward to use skeleton joint angles, joint angle velocities, and velocity of joints extracted from the raw motion for calculating a feature descriptor as applied in [72]. However, this approach requires the number of skeleton frame T should be equal for all of the obtained actions in order to create a fixed length of feature vector x_i. However, it is important to notice that collecting a fixed length of motion sequence is a challenging task since human behavior data vary from one behavior to another. Consequentially, the pose estimation module may produce different frame lengths for different actions. It is required that the feature descriptor phase should produce the fixed-length descriptors regardless of the length of the obtained skeleton frames. The covariance descriptor proposed by [73] can satisfy such a requirement and achieve higher accuracy compared to the other approaches [74].

The feature encoding process is started by calculating the covariance matrix $C(S)$ of the action A_i as described in Eq. 3.1. Here, $\overline{S}$ is the sample mean of S_i

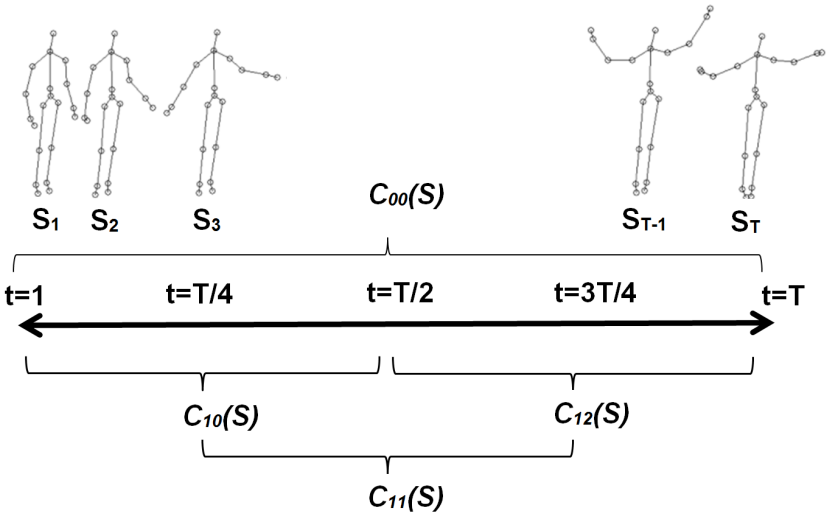


Figure 3.2: The hierarchy of overlapped covariance matrices.

computed over the time t and $^\intercal$ represents the transpose operator. S_i represents k joints of a human skeleton in 3D space. Consequently, $N = 3k$ elements are included in the vector S_i and the upper triangle of $C(S)$ contains $(N \times (N+1)/2)$ elements.

$$C(S) = \frac{1}{t-1} \sum_{i=1}^{t} (S_i - \overline{S})(S_i - \overline{S})^\intercal \tag{3.1}$$

It should be remarked that by simplify using the covariance matrix $C(S)$ described in Eq. 3.1, only the spatial features of action A_i could be presented. By combining several covariance matrices overlapped to each other over the time sequence, the spatial and temporal features of A_i could be determined. As demonstrated in Fig. 3.2, the matrix $C(S)$ at the level l would cover $t = T/2^l$ skeleton frames. Specially, $C_{00}(0)$ is the covariance matrix calculated at the level $l = 0$, it captures motion features of the entire action A_i including T frames. At the level 2, we calculated 3 smaller overlapping time windows, each of them would cover $T/2$ frames. The covariance matrix $C_{10}(S), C_{11}(S)$, and $C_{12}(S)$ is 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: $C_{00}(0)$, $C_{10}(S), C_{11}(S)$, and $C_{12}(S)$. 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 [73] and unsupervised learning tasks [75]. At the end of the action encoding phase, n action $A_1, A_2, ..., A_n$ are encoded into n fixed-length feature descriptor $x_1, x_2, ..., x_n$.

3.3.2 Training Phase

Self Organizing Map (SOM)

Given sets of n feature descriptors from the encoding phase 3.3.1, as an unsupervised learning approach without *a priori* knowledge of the number of clusters, Self Organizing Map (SOM) was implemented for the training phase in our previous work [76]. SOM was originally introduced by Kohonen [77], this approach creates a grid of neurons representing the distribution of the original data input. SOM ensures the topological property of the input data is preserved on the grip of training neurons [77]. Meaning that, if two patterns of human behavior are close to each other on the original motion space, the neurons representing that patterns would locate close to each other on the space of SOM neurons.

For the n input descriptors released from the action encoding phase, each of them $x_i = [x_{i1}, x_{i2}, ..., x_{id}]$ includes d-dimensional features. The training process is started by defining a SOM grid, including N $= p \times r$ neurons, each neuron represented by a prototype vector $m_p = [m_{p1}, m_{p2}, ..., m_{pd}]$. During the training process, an input sample x_i is picked up, then the wining neuron m_{wining} is determined by Eq. 3. m_{wining} is defined as the neuron has the shortest distance to the x_i.

$$||x_i - m_{wining}|| = min\{||x_i - m_i||\}, \quad 1 \leq i \leq N \tag{3.2}$$

The winning neuron m_{wining} is updated as illustrated in Eq. 3.3, allowing them to move closer to x_i with a highest intense comparing with the other neurons. Where $\alpha(t)$ defines the learning rate at the time t.

$$m_{wining} = m_{wining} + \alpha(t) \times (x_i - m_{wining}) \tag{3.3}$$

Not only the wining neuron m_{wining}, the neuron m_i which located near m_{wining}

(known as topological neighbor neurons) is also updated by Eq. 3.4, making them move closer to the input unit x_i.

$$m_i = m_i + \alpha(t) \times \phi(m_i, m_{wining}) \times (x_i - m_i) \tag{3.4}$$

Here $\phi(m_i, m_{wining})$ is the neighborhood kernel function, it indicates the intensity of the wining neuron m_{wining} affects its neighbor m_i. In the proposed framework, we used the Gaussian kernel function as illustrated in Eq. 3.5 where r_{wining} and r_i is the location of the winning neuron $m_{winning}$ and the neuron m_i on the grid map, respectively. It has been shown that by implementing the Gaussian kernel function, the global topological relationship could be better preserved on the grid of training neurons [78]. As the result, it encourages the training neurons to better reflect the distribution property of the original data input. This factor plays a crucial role in the next step: clustering the training neurons into different groups based on their similarities.

$$\phi(m_i, m_{wining}) = exp\left(-\frac{||r_{wining} - r_i||^2}{2\sigma^2(t)}\right) \tag{3.5}$$

Dynamic Cell Structure (DCS)

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 without corrupting the existing model. However, with the SOM network, the number of trained neurons must be fixed in advance, which makes this approach is inappropriate for incremental learning. As the number of input patterns incrementally increased, the network of training neurons should be equipped with the capability of extending its size in an incremental manner. To satisfy the requirement of incremental learning for scenarios of day-to-day interactions as well as ensuring topological preservation, we have employed the Dynamic Cell Structure (DCS) neural architecture [79] for the training phase [75]. DCS represents a family of artificial neural networks that could be applied in both supervised and unsupervised learning. It belongs to the

class of Topology Representing Networks which build perfectly topology preserving feature maps [80]. DCS inheres Kohonen type learning rule [77] for updating weight of neural vectors as applied in SOM, while using Hebbian learning rule [81] to dynamically update lateral connection structure (topology of the graph of neurons). As the result, DCS makes sure that topological properties are 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. The other approaches of growing neural networks by dynamic allocating the feature map are known as Growing Cell Structure (GCS)[82], Growing Neural Gas [83], and Grow When Required (GWR) [84]. In [85], GWR has been used as supervised learning to recognize the affective states of human bodily expression. Among techniques inspired by SOM, DCS works in a very similar way with GCS except for one essential difference: the lateral connections between neural units are not initially defined, instead, they are dynamically learned during the training phase [79] by Hebbian learning rule. DCS has been widely used for online learning purposes, such as the NASA first-generation Intelligent Flight Control System program [86].

On the DCS network, for the incoming input descriptor x_i, Eq. 3.6 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 [81] as described in Eq. 3.7, where ε is a forgetting constant and ϑ is a threshold for deleting lateral connection. It is also noted that the lateral connection C_{ij} between two neurons m_i and m_j is defined in the range from 0 to 1. $C_{ij} = 1$ if they are completely connected to each other and vice versa, $C_{ij} = 0$ if they are disconnected to each other. This lateral connection is always bidirectional and has symmetric weight.

$$\begin{aligned} ||x_i - m_{bmu}|| &\leq ||x_i - m_i||, & 1 \leq i \leq N \\ ||x_i - m_{second}|| &\leq ||x_i - m_i||, & 1 \leq i \neq bmu \leq N \end{aligned} \tag{3.6}$$

$$C_{ij}(t+1) = \begin{cases} 1 & ,(i = bmu) \wedge (j = second) \\ 0 & ,(i = bmu) \wedge (j \in \{N_i\} \setminus \{second\}) \\ & \quad \wedge (C_{ij} < \vartheta) \\ \varepsilon C_{ij}(t) & ,(i = bmu) \wedge (j \in \{N_i\} \setminus \{second\}) \\ & \quad \wedge (C_{ij} \geq \vartheta) \\ C_{ij}(t) & ,otherwise \end{cases} \tag{3.7}$$

Similar to SOM, DCS then updates their weight of neuron vectors by Kohonen learning rule [77] which makes them move closer to the current input as presented in Eq. 3.8. The neighbor neurons m_i is also updated, the intensity of changes is defined by Gaussian kernel function as illustrated in Eq. 3.5.

$$\begin{aligned} m_{bmu} &= m_{bmu} + \eta(t)(x_i - m_{bmu}) \\ m_i &= m_i + \eta(t)\phi(m_i, m_{bmu})(x_i - m_i) \end{aligned} \tag{3.8}$$

The resource value τ_{bmu} of the closest neuron m_{bmu} is updated by Eq. 3.9. The new neuron unit m_{new} could be added into the network and locates between neurons with the largest and second-largest resource value. The training phase is finished by decreasing the resource value τ_i of all neuron units, as described in Eq. 3.11, where λ is defined as the decreasing rate.

$$\tau_{bmu} = \tau_{bmu} + ||x_i - m_{bmu}||^2 \tag{3.9}$$

$$E_q = ||x_i - m_{bmu}||^2 \tag{3.10}$$

$$\tau_i = \lambda \tau_i \tag{3.11}$$

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.

3.3.3 Clustering Phase

In the previous section, a grid of neurons could be trained by SOM (as a batch learning version) or DCS (as an incremental learning version) approach. As the proposed framework presented in Fig. 3.1, in the clustering phase, a grid of training neurons will be separated into different groups based on their similar features. Several approaches have been suggested such as agglomerative clustering or k-means algorithm [87] [88]. By using k-means for clustering neurons, this involved making several k-means clustering trials with different values of k [87] and the best clustering should minimize the value of the Davies-Bouldin index [89]. However, the minimum value of the Davies-Bouldin index was not always indicating the appropriate number of clusters. In [90], the authors utilized a distance matrix to identify cluster centers from a grid of training neurons. Then, the other neurons are assigned to the corresponding clusters based on distances between them and the identified centers. It is noticed that a distance matrix indicates distances between each of the neurons and their neighbors. The distance matrix based clustering takes the most advantage of SOM - topological preservation. As the result, distances between neighboring neurons are approximately proportional to the distribution of the original data [90]. Fig. 3.3 presents an example case of using distance matrix visualization technique, named the unified distance matrix (U-matrix), for visualizing distances between training neurons [76]. The following part will explain the distance matrix based approach in detail.

From a grid of training neurons provided from the training phase, the clustering phase is started by identifying local minima (representative local neurons) of the distance matrix by Eq. 3.12. The function $f(m_i, N_i) = median\{||m_i - m_j||\}$ presents the median distance between neuron m_i and its neighboring neurons m_j.

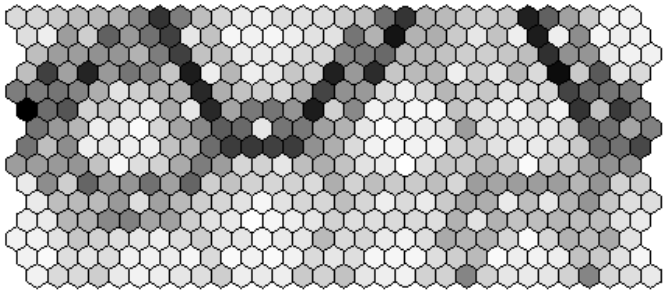


Figure 3.3: Visualization a grid of training neurons by U-Matrix, dark colors indicate larger distances between neuron units and their neighbors.

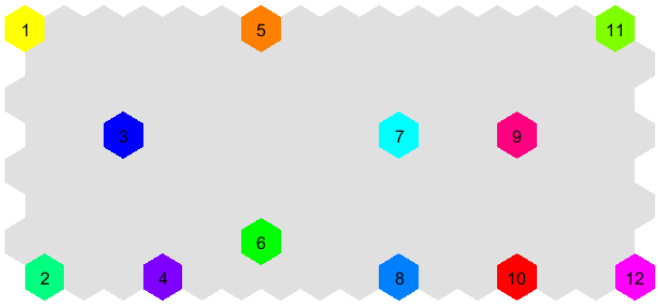


Figure 3.4: The detected local minima neurons (colored hexagons) and the unassigned ones (gray hexagons) on the training grid.

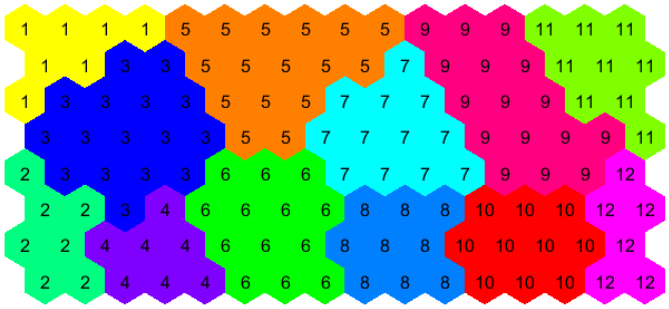


Figure 3.5: The unassigned neurons are assigned into appropriate clusters based on the distance between them to the local minima neurons

Fig. 3.4 shows a set of identified local minima m_i of the grid of training neurons.

$$f(m_i, N_i) \leq f(m_j, N_j) \quad \forall j \in N_i \tag{3.12}$$

After defining local minima neurons, each local minima represents a cluster. The unassigned neurons are put into corresponding clusters taking into account

distances between them and the closest local minima neurons, as shown in Fig. 3.5. Since each neuron unit m_i creates a Voronoi region on the original space of feature descriptors x given by Eq. 3.13. As the result, each neuron m_i and its corresponding input data x_i is defined by the Best Matching Unit function given by Eq. 3.14. The descriptor x_i belongs to the same cluster as its corresponding neurons m_i.

$$V_i = \{x| \quad ||x - m_i|| \leq ||x - m_j|| \quad \forall j \neq i\} \tag{3.13}$$

$$||x_i - m_i|| = min\{||x_i - m||\} \tag{3.14}$$

3.3.4 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 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. 3.15. Now the descriptor x_{rep} is the one located closest to the center μ of the $Cluster_i$. 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}.

$$||x_{rep} - \mu|| \leq ||x - \mu||, \quad \forall x \in Cluster_i, \tag{3.15}$$

Overall, for a new input action A_i obtained, the designed framework using the

Algorithm 1 The proposed framework processes a newly observed action A_i.

Input: observed action A_i, current network $curI$,
network parameters ϵ, ϑ, η, ϕ, λ;

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

DCS approach 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.

Chapter 4

Generation of Communicative Gestures using Conditional Generative Adversarial Network

In this chapter, we address the problem of generating communicative gestures supporting for semantic contents of communicators' speech. In section 4.1, we provide a review of previous studies in generating robots' co-speech gestures inspired by the rule-based approach. It is followed by discussions about recent studies based on the data-driven approach. Finally, the proposed approach as well as the framework architecture are presented in detail. Noticed that experiments conducted to evaluate this framework can be found in section 6.3 of chapter 6.

4.1 Related Works

During social human-robot interaction, communicative gestures provide robots capability of using bodily expressions for emphasizing their speech or describing something that they are talking about. This non-verbal channel helps robots' intentions are more understandable to interacting partners. Especially for the robot without dedicated facial articulation such as Pepper or NAO robot, communicative gestures support contexts of robots' speech that can be transmitted to humans in a facile and transparent manner [91]. Understanding the importance of co-speech

gestures in social robots, there has been an increasing interest in the creation of robots' actions synthesized with verbal contents of robots' speech. The studies in this domain could be broadly categorized into two groups: rule-based approach and data-driven approach.

It should be emphasized that the majority of existing works on generating communicative robot gestures rely on the ruled-based approach. In [22], the authors proposed Behavior Expression Animation Toolkit (BEAT) which receives the input text to be spoken and releases the non-verbal behaviors. In the BEAT toolkit, the mapping from text to gesture is based on a set of rules derived from the state of the art in non-verbal conversational behavior researches. Although this approach can produce various gestures, the basic motions must be designed manually. The model proposed in [92] accepts both lexical contents of utterance and audio signals as the inputs to generate the non-verbal behaviors for virtual agents. Similar to the BEAT toolkit, the basic behaviors must be designed in advance. Recently, several advanced social robots such as RoboThespian, Nao, and Pepper have become capable of making the communicative gestures synchronized with their speech, but their gestures are handcrafted by animation experts in order to ensure the familiarity and human-likeness of the gesture.

Although the handcrafted gestures provide the familiarity and human-likeness of the robot motions, this approach only allows robots to produce their communicative behaviors in the pre-designed scenarios. Moreover, the generated gestures are limited to a set of rules. It should be reported that social robots need to be capable of interacting with different types of users in a personal way by adapting and learning its behaviors throughout its lifetime [93]. Thus, social robots should be endowed with the capability of learning social skills from perceived human behaviors. This idea resembles how infants learn social behaviors from their parents that we have described in chapter 2. Inspired by infant social developments [51], several studies have been conducted for producing facial [45] and bodily expressions [76] for social robots. Taking into account theories of human emotion [94], it is well known that emotion could be categorized into several basic groups (*happy, sad, surprised, disgusted, angry, fearful*). Thus, each of the emotions could be treated in an appropriate manner as the model for generating emotional gestures described in chapter 3. On the other hand, communicative behaviors are more complex and

they require a highly sophisticated model. In other words, to generate co-speech gestures, relations between behaviors and corresponding natural language context need to be addressed in a variety of communication topics. Recently, this approach has received increasing attention in the social robotics domain. In [95], the authors proposed the $3D$ pose generation model utilizing the recurrent neural network. The model receives speech audio features and/or text input to generate the gestures corresponding to certain specific words. The generated upper body motions are represented by the human joint coordinates. Afterward, they are converted to the target robot joint angles. In [96], the authors presented a framework for speech-driven gesture generation. The network is designed based on auto encoder-decoder. The framework receives audio features (MFCCs, spectrograms, prosodic) as inputs and produces an output body motion sequence. Similarly, a speech-driven model for facial motion generation can be found in [97], this framework is built upon bidirectional long-short term memory. On the other hand, the authors [98] suggested the co-speech gesture generation framework which receives the raw text input. Through the encoder and decoder phases, the upper body poses are released. Then, the generated motions are re-targeted to the Nao humanoid robot. Although various co-speech gestures could be generated by the authors' proposed approach, it is suggested that the model is not able to learn iconic or metaphoric gestures in an efficient manner [99]. In [100], the bidirectional relation between the human whole-body motion and natural language was investigated. The authors demonstrated the capability of their proposed framework to generate text descriptions for a variety of human body motions. Vice versa, with the text description input, the model produces the gestures displayed on the Master Motion Map (MMM) model. However, generated actions are defined in joint space with respect to the MMM joint configuration, it is difficult to utilize this approach on the other robots whose kinematic structures are different from the MMM framework. Recently, Generative Adversarial Network (GAN) has received considerable attention in a variety of domains, especially for image generation tasks [101]. To the best of our knowledge, Text2Action [1] is the first paper using a GAN framework for generating robots' co-speech actions synthesized with the input context. It is constructed based on a sequence to sequence network. Different from Text2Action, our generative framework is built upon convolutional

neural network (CNN) which has been widely used in many research contexts such as image [102, 103], video [104], and audio generation [105]. Inspiring by those successes, our research investigates the convolution operation toward the autonomous generation of communicative actions.

4.2 Research Approach

It is noticed that different aspects could be considered when generating gestures synthesized with communicators' speech. This proposed approach focuses on generating co-speech gestures supporting the concrete or abstract contents of users' speech. Taking into account theories of human communicative gestures as discussed in chapter 2, those gestures are known as *iconic*, and *metaphoric* gestures. Other types of gestures such as *deictic* or *beat* fall outside scope of our work. The proposed approach uses GAN to learn relations between human communicative gestures and semantic contents of their speech. Although GAN has received considerable attention across different disciplines. Generating robot motions with GANs, however, is seldom explored [106]. Our research aims at extending the application of GAN for generating social robots' non-verbal actions when synthesizing their verbal content of speech. In a GAN network, the Generator and Discriminator networks are simultaneously trained and updated. The Generator tries to create the samples imitating the training data distribution, while the Discriminator tries to distinguish between generated samples and real data of the training set. Consequentially, Generator G and Discriminator D play a min-max game as given in Eq. 4.1. In the conventional GAN, the network receives a noise vector sampled from a prior distribution to generated fake data. Taking into account our research topic, the generated data would be robots' co-speech actions. In order to ensure that robots' generated gestures are highly connected to their verbal content of speech, the relation between robots' actions and synthesizing text should be carefully considered. Let us assume that the sentences which are uttered by a robot are the determining factor for generating the robot's gestures. This connection can be taken into account using Conditional Generative Adversarial Network (CGAN) approach [107], an extension of GAN with additional input condition c to control to control output data. As the result, the objective function

of the min-max game between G and D would be as Eq. 4.2. In our proposed framework, CGAN is built upon CNN to generate communicative gestures when synthesizing the verbal content of speech. The following section will explain our designed framework in detail.

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}}\left[logD(x)\right] + \mathbb{E}_{z\sim p_z}\left[log(1-D(G(z)))\right] \tag{4.1}$$

$$\min_G \max_D V(D,G) = \mathbb{E}_{x,c\sim p_{data}}\left[logD(x,c)\right] + \mathbb{E}_{c\sim p_{data},z\sim p_z}\left[log(1-D(G(z,c),c))\right] \tag{4.2}$$

4.3 Framework Architecture

Fig. 4.1 presents the proposed framework for generating robot's co-speech actions. In the training phase, $a_r = [S_1, S_2, S_3, ..., S_T]$ ($a_r \in \mathbb{R}^{3\times 8\times T}$) denotes a real action from the training data that contains a sequence of skeleton frames $S \in \mathbb{R}^{3\times 8}$ performed over a period of time T. As shown in Fig. 4.4, S consists of 8 joints defined in 3D space. Through the Action Encoder, a_r is encoded to an action matrix $x_r \in \mathbb{R}^{3\times 16\times T}$. On the other hand, $d = [w_1, w_2, w_3, ..., w_k]$ is a natural language sentence composed of k words to describe the action a_r. It is started by feeding the description d to the Embedding Description network. The output e is concatenated with the noise vector z sampled from the normal distribution, and they are fed to the Generator network. The purpose of the Generator G is to generate the fake action matrix $x_f \in \mathbb{R}^{3\times 16\times T}$ as much realistic as possible to beat the Discriminator D while D tries to differentiate between x_r and x_f taking into account the embedding vector e. Once the training process is completed, the generated action matrix x_f, synthesized with text description d, is decoded to $a_f \in \mathbb{R}^{3\times 8\times T}$. Through the Transformation model, the action a_f, defined in 3D Cartesian space, is transformed into the target robot's motion space represented by joint angles.

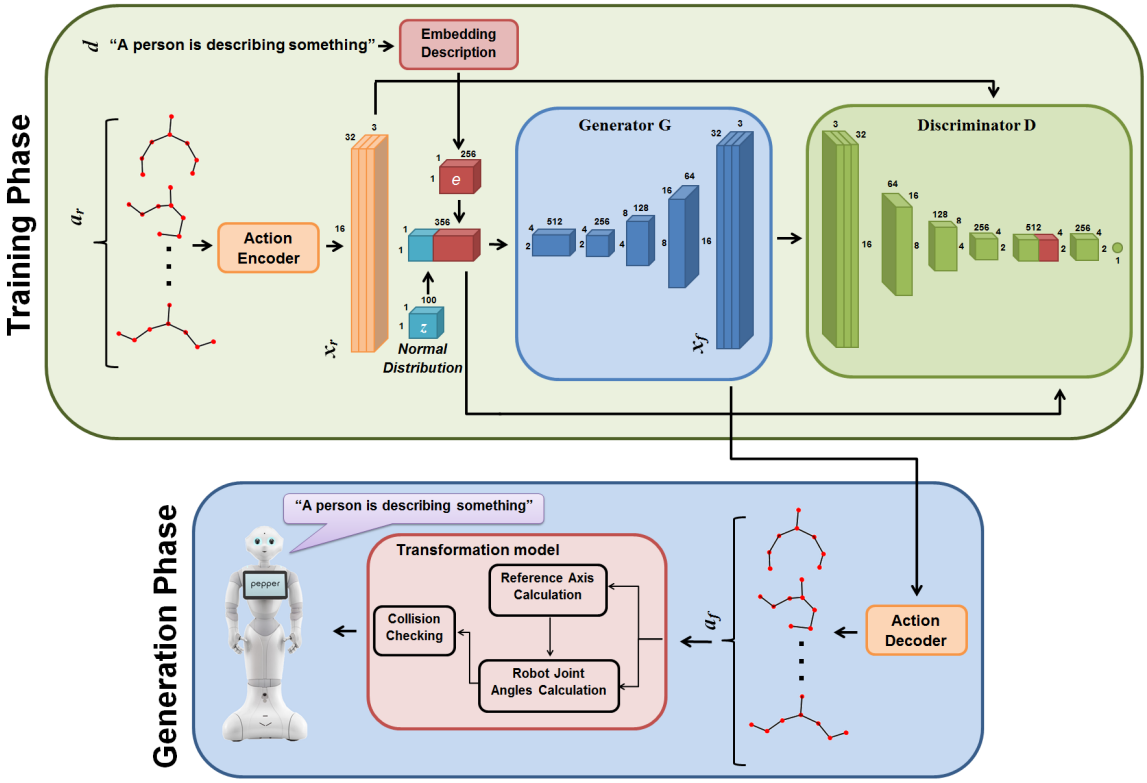


Figure 4.1: The proposed framework for generating fake action a_f synthesized with text input d. Through the transformation model, generated action a_f is transformed into the target robot motion, being the robot's social gesture.

4.3.1 Embedding Descriptor

In order to encode the input description into the fixed-length embedding vector e, which efficiently captures the meaning of the whole sentence, $d = [w_1, w_2, w_3, ..., w_k]$ is fed into the Embedding Description. Here, we use the encoder phase of the skip-thoughts model [108]. The output vectors from this model effectively represent the semantics and syntax of the sentence to be encoded [108].

$$h_k = (1 - z_k) \odot h_{k-1} + z_k \odot tanh(W c_k + U(r_k \odot h_{k-1})) \tag{4.3}$$

The hidden layer h_k represents the sequence of words $\{w_1, ..., w_k\}$. h_k is calculated by Eq. 4.3, where c_k is the word embedding of w_k, W, U are the weight matrices, $\odot$ denotes a component-wise product, z_k and r_k represent the update

gate and reset gate of Gated Recurrent Unit [109], respectively. The hidden state h_k captures the meaning of the whole sentence d, this value is then compressed into a smaller dimensional vector e before being fed into the Generator and Discriminator model.

4.3.2 Action Encoder and Decoder

Action Encoder

Convolutional Neural Network (CNN) has a natural ability to learn representation from 2D matrices [110]. Human actions, defined as a sequence of skeleton frames, could be represented as 2D matrices containing three channels representing x, y, z coordinates, respectively. On each channel, the horizontal axis shows the time sequence of skeleton frames, while the vertical axis represents the spatial distribution of joints at a certain timestamp. Then, CNN based approach is utilized to jointly capture spatial and temporal information of actions [111, 110, 112]. It should be emphasized that the chain order of joints in the vertical axis affects the spatial information represented in the action matrix x_f. To efficiently capture spatial relations of the adjacent joints of the action a_r, the Action Encoder puts its relative joints near each other. With this representation, by feeding the input a_r to the Action Encoder, the encoded matrix x_r is released. This can be seen in Fig. 4.2. Specifically, on each channel $c \in \{x, y, z\}$ of the matrix x_r, the horizontal axis covers the time sequence T of the action a_r, while the vertical axis is a sequence of joints in a given order $I = [1, 0, 1, 2, 3, 4, 3, 2, 1, 1, 5, 6, 7, 6, 5]$ ($I \in \mathbb{R}^{16}$) at a certain timestamp. Thus, instead of feeding the raw input a_r to the D network, the Action Encoder allows the spatial-temporal information of the action a_r to be presented as the action matrix x_r.

Action Decoder

In order to decode the action matrix x_f to the action a_f as displayed in Fig. 4.1, our designed Action Decoder calculates the joint value $j_{c,m,t}$ of the action a_f over the time sequence as shown in Eq. 4.4 and Eq. 4.5. This calculation allows that $j_{c,m,t}$ is defined based on the average values of its distribution on x_f. Here, $j_{c,m,t}$

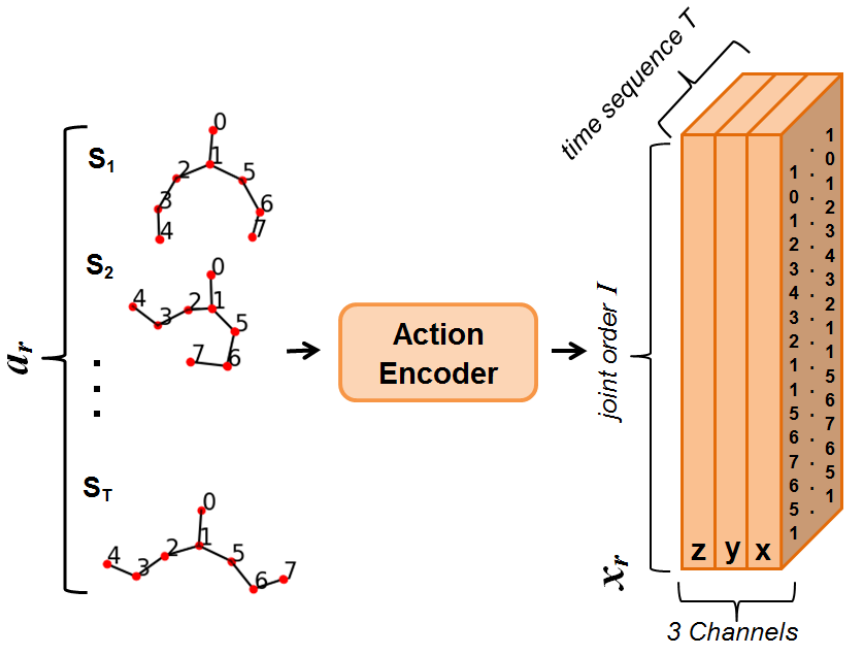


Figure 4.2: Action Encoder encodes the raw action a_r to the action matrix x_r

denotes the value of joint index m ($m = [0, 7]$), on the dimension c ($c \in \{x, y, z\}$), at the time stamp t ($t \in [1, T]$), and $n(m)$ is the number of times the joint index m in the order I.

$$j_{c,m,t} = \frac{1}{n(m)} \sum_{p=1}^{16} x_f(c, p, t)\delta(p, m, I) \tag{4.4}$$

$$\delta(p, m, I) = \begin{cases} 1 & I(p) = m \\ 0 & I(p) \neq m \end{cases} \tag{4.5}$$

4.3.3 Generator

The proposed model is based on the transposed convolutional network which has been shown to be useful in many different research contexts such as image generation [102, 103], video generation [104], and audio generation [105]. Initially, the noise vector z is sampled from the Normal distribution $N(0, 1)$. It is concatenated with the vector e, encoded from Embedding Description, before being fed to the

G network. As presented in Fig. 4.1, G is designed with a fully connected layer to reshape the input vector and followed by four fractionally-strided convolutions to up-sample the data to an output target x_f. On each layer, batch normalization is utilized for stabilizing the learning process. This operation normalizes the input to each unit to have zero mean and unit variance. The output values are followed by the Rectified Linear Unit (ReLU) activation [113] except for the last layer. Here, the *tanh* activation function is used before producing x_f.

4.3.4 Discriminator

Discriminator D is designed with five convolutional layers similar to the architecture of G. D receives either x_r from training data or x_f from G as an input. At the fourth layer, the embedding vector e is concatenated with the output of the convolutional layer. Here, the embedding e provides conditional information to D in order to evaluate whether the input action satisfies this condition or not. At the last layer, the results are passed into a sigmoid function to produce an output probability.

The training process is summarized in Algorithm 2. The vector e provides conditional information to the G network in order to generate the action matrix x_f, synthesized with the action description d. The aim of the Generator is to fool the Discriminator. Thus, the Generator is trained to maximize the output probability y_f. Conversely, D is trained to differentiate between x_r and x_f based on (1) the human-likeness of the action, and (2) the synthesis of an action and its corresponding description. It should be remarked that the second point plays an essential role, allowing the generated action to effectively express the meaning of input description. To endow D with the capability of evaluating this synthesis, D is trained to maximize the output probability y_r when receiving a pair of real action input x_r and embedding vector e. On the other hand, given a pair of input x_f and e, the Discriminator is trained to minimize the output probability y_f. From the training data, we also collect the miss-matching description $\hat{d}$, which incorrectly describes the action x_r. When feeding a pair of the real action x_r and $\hat{e}$ to the D network, the Discriminator is trained to minimize the output y_m, implying that x_r does not synthesize $\hat{d}$. The binary cross-entropy is applied to compute the miss-

classification error L_D, L_G of the network D and G, respectively. The parameter of D is updated while keeping the parameters of G constant. Then, the parameters of G are adjusted to optimize the error L_G while keeping network D unchanged.

Algorithm 2 The proposed algorithm for training the Generator G and the Discriminator D.

Input: real action a_r, matching description d, miss-matching description $\hat{d}$, training batch steps S.

for s=0 to S **do**
  $x_r \leftarrow \text{ActionEncoder}(a_r)$;
  $e \leftarrow \text{EmbeddingDescription}(d)$;
  $\hat{e} \leftarrow \text{EmbeddingDescription}(\hat{d})$;
  $z \leftarrow \text{N}(0, 1)$;
  $x_f \leftarrow \text{G}(z, e)$;
  $y_r \leftarrow \text{D}(x_r, e)$;
  $y_f \leftarrow \text{D}(x_f, e)$;
  $y_m \leftarrow \text{D}(x_r, \hat{e})$;
  $L_D \leftarrow log(y_r) + log(1 - y_m) + log(1 - y_f)$;
  $D \leftarrow D - \alpha(\partial L_D / \partial D)$; {Update Discriminator}
  $L_G \leftarrow log(y_f)$;
  $G \leftarrow G - \alpha(\partial L_G / \partial G)$; {Update Generator}
end for

Chapter 5

Transforming Generated Human-like Gestures into the Target Social Robot

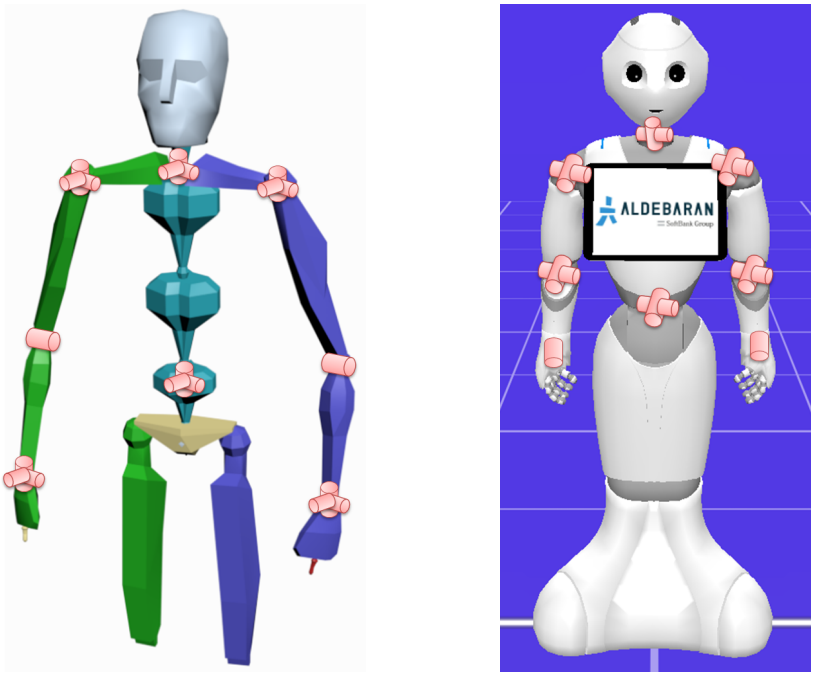
In chapter 3 and 4, the frameworks for generating emotional and communicative gestures were presented. However, they are designed to learn from human behaviors, as the results, generated gestures are defined in human motion space. In this chapter, we will explain the proposed approach to transform generated human-like gestures into the target Pepper humanoid robot, taking into account the robot's physical constraints. Experiments conducted to verify this framework as a standalone function can be found 6.1 of chapter 6. Indeed, the integration of this model and the ones illustrated in chapter 3 and 4 could be found in section 6.2 and 6.3, respectively.

5.1 Related Works

Recently, social robots such as NAO, Pepper, and RoboThespian have become capable of performing human-like gestures towards enhancing the quality of human-robot interaction. However, it is always in case that such gestures are programmed and implemented in advance by animation or robotics experts. One of the limitations of this approach is that it does not allows non-expert users to teach robots

new social behaviors supporting typical scenarios of interactions. There is an increasing need that robots should be able to re-configure their interacting gestures adapting to interacting environments towards increasing users' empathy and engagement of social interaction. To archive that target, imitation learning is a well-established concept. It appears as a promising approach for teaching robots new motions through demonstration [114, 13], robots then generate human-like movements similar to demonstrators' actions.

A common approach for imitation learning is sensing human actions as 3D motion data using optical marker sensors (Vicon, Phase Space) or markerless sensors (Kinect Microsoft, Asus Xtion) (further information can be found in Appendix). Then, through a designed transformation model for solving the problem of inverse kinematic, motion data of demonstrators is converted to a set of joint angles representing robots' gestures. In [115], the authors proposed a system running in ROS environment [116] to teleoperate the NAO robot's arms. A human motion is captured by the Kinect sensor. Through a transformation model based on the geometric inverse-kinematic, a human motion is transformed into the target robot. It is noticed that with legged robots, in addition to the transformation process for calculating imitated robot's joint angles, positions of Center of Mass (COM) [117] should be taken into account to maintain the robot's balance. This problem has been addressed when generating imitated gestures for the NAO robot in [118]. A similar approach can be found in [119] for imitating human motions during an on-line demonstration. Through a transformation model for solving inverse kinematic, the upper body motion of the Darwin-OP humanoid robot is generated while the balance of the robot is guaranteed. In their proposed approach, this problem is solved by optimizing the robot's motion around COM. On the other hand, rather than designing a particular transformation model for a specific robot's configuration, the authors [12] introduces a generic transformation model, allowing human gestures could be imitated by different robot platforms. The authors demonstrated their proposed approach on a public human bodily expression dataset [14]. Through the generic model, human affective gestures are transformed into different robot platforms such as ASIMO, Justin, and NAO robot while meanings of human bodily expression remained unchanged. In contrast with the conventional approach, where the problem of inverse kinematic is taken into



(a) Human joint configuration.

(b) Pepper robot joint configurartion.

Figure 5.1: The availability of DoFs on the upper body of the human model and the target robot model.

account to determine the mapping between human joint coordinates and angular position of the target robot's actuators, recent advantages of deep neuron network provide an alternative approach to tackle this issue. In [120], the authors firstly collected paired synchronized movements capturing both human motion data and the target robot's actuator data. By training a feed-forward neural network for each Degree of Freedom (DoF) on the robot, the relations between human joint coordinates and the corresponding robot's DoF are detected. Similarly, by utilizing the machine learning approach, a mechanism for human whole-body imitation was introduced in [29]. Finally, it should be noticed that because of differences in joint configurations between humans and robots, self-collision may exist on generated robots' gestures. By equipping the imitation learning model capability of self-collision avoidance, the Tangy robot [13] is capable of producing collision-free movements imitating the demonstrator's actions.

5.2 Research Approach

Fig. 5.1 shows the differences in joint configurations between a human model and the Pepper robot. It can be seen that the DoFs of the target robot are limited compared to the human model. As the result, our designed transformation model presented in Fig. 5.2 converts human actions into a set of joint angles displayed on the target robot, taking into account the robot's kinematic structure. For calculating the Pepper robot's joint angles, the solutions to the inverse kinematic problem are computed based on geometric algebra. This approach has been widely used in previous studies [116, 119, 13] mentioned above. It is also noticed that there are significant differences in the lower body between humans and the target robot. As the result, the proposed approach focuses on the imitation of the robot's upper body including the movements of *hip, head, shoulder, elbow*, and *wrist* on both the left and right sides. The following section will detail our proposed framework.

5.3 Framework Architecture

Fig. 5.2 illustrates the architecture of the proposed approach. In the designed model, the robot joint angle calculation phase receives the human joint vectors Left/Right Knee (l_k, r_k), Left/Right Hip (l_hi, r_hi), Central Hip (c_hi), Torso (tor), Neck ($neck$), Left/Right Head (l_he, r_he),Left/Right Shoulder (l_s, r_s), Left/Right Elbow (l_e, r_e), Left/Right Hand (l_h, r_h), and the axes (x_ref, y_ref, z_ref) computed from the reference axis calculation phase. A set of robot's angles released from the joint angle calculation phase are passed through the collision checking phase before inputting to the robot's actuators.

5.3.1 Reference Axis Calculation

During social interaction, it is common that robots perform nonverbal communicative behaviors such as head motions 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, it is necessary that the reference axes should be independent of the camera configuration.

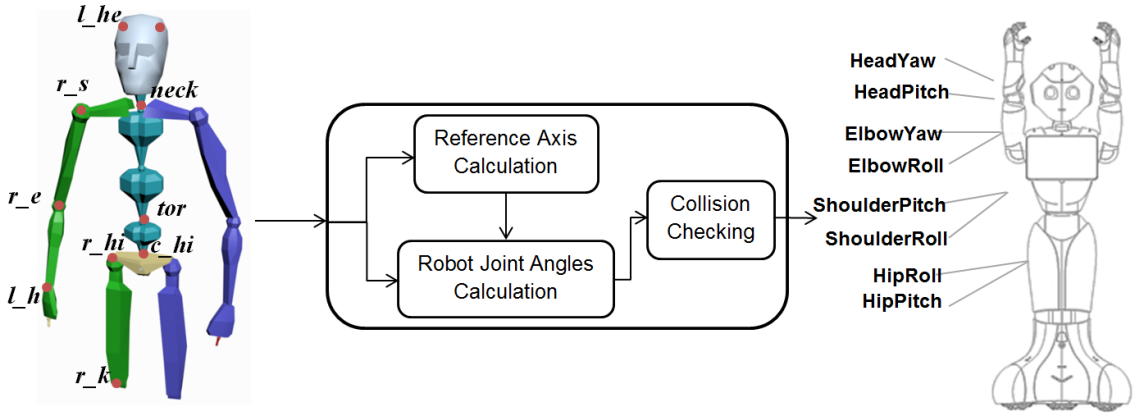


Figure 5.2: Transformation of human joint positions into the Pepper robot's joint angles

Thus, we used the reference axes calculated by Eq. 5.1, 5.2, and 5.3 to describe the orientation of the estimated pose. The calculated axes x_ref, y_ref, z_ref are combined with the human motion data input for calculating the robot joint angles.

$$\overrightarrow{z_{ref}} = (\overrightarrow{r\_hi} - \overrightarrow{l\_hi}) \times (\overrightarrow{r\_hi} - \overrightarrow{tor}) \tag{5.1}$$

$$\overrightarrow{x_{ref}} = \overrightarrow{r\_hi} - \overrightarrow{l\_hi} \tag{5.2}$$

$$\overrightarrow{y_{ref}} = (\overrightarrow{z_{ref}} \times \overrightarrow{x_{ref}}) \tag{5.3}$$

5.3.2 Joint Angle Calculation

The join angle calculation phase receives a set of human motion data and the reference axes x_ref, y_ref, z_ref as the inputs. We applied the geometric algebra approach for solving the problem of inverse kinematic. In the end, this calculation

phase releases a set of joint angles Roll (α), Pitch (β), Yaw (γ) corresponding to the availability of DOFs of the Pepper robot kinematic structure. Depending on the robot's home configuration different from that of the human joints, an offset value is added to the calculated joint.

It is started determining movements of the robot's hip. In the target robot model shown in Fig. 5.1b, α_{Hip} controls the side-to-side movement while up-and-down motion is manipulated by β_{Hip}. Eq. 5.4 and Eq. 5.5 are used to determine that two DoFs.

$$\alpha_{Hip} = -arccos\left(\frac{(\overrightarrow{l\_s} - \overrightarrow{r\_s}) \cdot \overrightarrow{y_{ref}}}{|\overrightarrow{l\_s} - \overrightarrow{r\_s}| \cdot |\overrightarrow{y_{ref}}|}\right) + \frac{\pi}{2} \tag{5.4}$$

$$\begin{gathered} \overrightarrow{b} = \overrightarrow{c\_hi} - \left(\frac{\overrightarrow{r\_k} + \overrightarrow{l\_k}}{2}\right) \\ \beta_{Hip} = -arccos\left(\frac{\overrightarrow{b} \cdot \overrightarrow{z_{ref}}}{|\overrightarrow{b}| \cdot |\overrightarrow{z_{ref}}|}\right) + \frac{\pi}{2} \end{gathered} \tag{5.5}$$

Concerning the robot's upper right arm, the side to side movement is illustrated by $\alpha_{RightShoulder}$. This value could be determined based on coordinates of the right shoulder and the right elbow as described in Eq. 5.6. Vice versa, $\beta_{RightShoulder}$ describes the up-and-down movement of the robot upper arm. As shown in Eq. 5.7, $\beta_{RightShoulder}$ is calculated by taking into account two neighboring vectors r_s and r_e, then combining with the reference axis y_{ref}.

$$\alpha_{RightShoulder} = -arccos\left(\frac{(\overrightarrow{r\_e} - \overrightarrow{r\_s}) \cdot (\overrightarrow{l\_s} - \overrightarrow{r\_s})}{|\overrightarrow{r\_e} - \overrightarrow{r\_s}| \cdot |\overrightarrow{l\_s} - \overrightarrow{r\_s}|}\right) + \frac{\pi}{2} \tag{5.6}$$

$$\beta_{RightShoulder} = arccos\left(\frac{\overrightarrow{y\_ref} \cdot (\overrightarrow{r\_s} - \overrightarrow{r\_e})}{|\overrightarrow{y\_ref}| \cdot |\overrightarrow{r\_s} - \overrightarrow{r\_e}|}\right) - \frac{\pi}{2} \tag{5.7}$$

The angle $\alpha_{RightElbow}$ is created by the two links, upper arm, and forearm. Thus, this angle could be defined by the dot product between that two neighboring links

as described in Eq. 5.8. On the other hand, $\gamma_{RightElbow}$ is created by rotation movement of forearm around the upper arm, this value is calculated as given by Eq. 5.9.

$$\alpha_{RightElbow} = -arccos\left(\frac{(\overrightarrow{r\_s} - \overrightarrow{r\_e}) \cdot (\overrightarrow{r\_h} - \overrightarrow{r\_e})}{|\overrightarrow{r\_s} - \overrightarrow{r\_e}| \cdot |\overrightarrow{r\_h} - \overrightarrow{r\_e}|}\right) + \frac{\pi}{2} \tag{5.8}$$

$$\overrightarrow{c} = R_z(\alpha_{RightShoulder}) \cdot R_y(\beta_{RightShoulder}) \cdot \frac{\overrightarrow{z_{ref}}}{|\overrightarrow{z_{ref}}|}$$

$$\overrightarrow{d} = \frac{(\overrightarrow{r\_s} - \overrightarrow{r\_e}) \times (\overrightarrow{r\_h} - \overrightarrow{r\_e})}{|\overrightarrow{r\_s} - \overrightarrow{r\_e}| \times |\overrightarrow{r\_h} - \overrightarrow{r\_e}|} \tag{5.9}$$

$$\gamma_{RightElbow} = -arccos\left(\frac{\overrightarrow{c} \cdot \overrightarrow{d}}{|\overrightarrow{c}| \cdot |\overrightarrow{d}|}\right) + \frac{\pi}{2}$$

Additionally, with the human motion capture Front Head (fr_he), Back Head (ba_he), Left/Right Wrist Near Thumb (lw_ra, rw_ra), Left/Right Wrist Opposite Thumb (lw_rb, rw_rb) are given, the robot's joint angle β_{Head}, γ_{Head}, and $\gamma_{RightWrist}$ can be calculated as the following:

$$\overrightarrow{d} = \left(\frac{\overrightarrow{l\_he} + \overrightarrow{r\_he}}{2}\right) \tag{5.10}$$

$$\beta_{Head} = arccos\left(\frac{(\overrightarrow{tor} - \overrightarrow{c\_h}) \cdot (\overrightarrow{d} - \overrightarrow{neck})}{|(\overrightarrow{tor} - \overrightarrow{c\_h})| \cdot |(\overrightarrow{d} - \overrightarrow{neck})|}\right)$$

$$\gamma_{Head} = arccos\left(\frac{(\overrightarrow{r\_s} - \overrightarrow{l\_s}) \cdot (\overrightarrow{fr\_he} - \overrightarrow{ba\_he})}{|(\overrightarrow{r\_s} - \overrightarrow{l\_s})| \cdot |(\overrightarrow{fr\_he} - \overrightarrow{ba\_he})|}\right) \tag{5.11}$$

$$\overrightarrow{g} = \overrightarrow{rw\_ra} - \overrightarrow{rw\_rb}$$

$$\gamma_{RightWrist} = -arccos\left(\frac{(\overrightarrow{-d}) \cdot \overrightarrow{g}}{|(\overrightarrow{-d})| \cdot |\overrightarrow{g}|}\right) + \frac{\pi}{2} \tag{5.12}$$

Similarly, the following equations are given to show the computation of joint angles, $\alpha_{LeftShoulder}$, $\beta_{LeftShoulder}$, $\alpha_{LeftElbow}$, $\gamma_{LeftElbow}$, and $\gamma_{LeftWrist}$, on the left side of the Pepper robot:

$$\alpha_{LeftShoulder} = -arccos\left(\frac{(\overrightarrow{l\_e} - \overrightarrow{l\_s}) \cdot (\overrightarrow{r\_s} - \overrightarrow{l\_s})}{|\overrightarrow{l\_e} - \overrightarrow{l\_s}| \cdot |\overrightarrow{r\_s} - \overrightarrow{l\_s}|}\right) - \frac{\pi}{2} \tag{5.13}$$

$$\beta_{LeftShoulder} = arccos\left(\frac{\overrightarrow{y\_ref} \cdot (\overrightarrow{l\_s} - \overrightarrow{l\_e})}{|\overrightarrow{y\_ref}| \cdot |\overrightarrow{l\_s} - \overrightarrow{l\_e}|}\right) - \frac{\pi}{2} \tag{5.14}$$

$$\alpha_{LeftElbow} = arccos\left(\frac{(\overrightarrow{l\_s} - \overrightarrow{l\_e}) \cdot (\overrightarrow{l\_h} - \overrightarrow{l\_e})}{|\overrightarrow{l\_s} - \overrightarrow{l\_e}| \cdot |\overrightarrow{l\_h} - \overrightarrow{l\_e}|}\right) - \frac{\pi}{2} \tag{5.15}$$

$$\overrightarrow{k} = R_z(\alpha_{LeftShoulder}) \cdot R_y(\beta_{LeftShoulder}) \cdot \frac{\overrightarrow{z_{ref}}}{|\overrightarrow{z_{ref}}|}$$

$$\overrightarrow{m} = \frac{(\overrightarrow{l\_s} - \overrightarrow{l\_e}) \times (\overrightarrow{l\_h} - \overrightarrow{l\_e})}{|\overrightarrow{l\_s} - \overrightarrow{l\_e}| \times |\overrightarrow{l\_h} - \overrightarrow{l\_e}|} \tag{5.16}$$

$$\gamma_{LeftElbow} = arcos\left(\frac{\overrightarrow{k} \cdot \overrightarrow{m}}{|\overrightarrow{k}| \cdot |\overrightarrow{m}|}\right) - \frac{\pi}{2}$$

$$n = \overrightarrow{lw\_ra} - \overrightarrow{lw\_rb}$$

$$\gamma_{LeftWrist} = arcos\left(\frac{(\overrightarrow{-m}) \cdot \overrightarrow{n}}{|(\overrightarrow{-m})| \cdot |\overrightarrow{n}|}\right) - \frac{\pi}{2} \tag{5.17}$$

Due to the differences in the lower body between human and the Pepper robot, the imitation of knee movement is ignored on the Pepper robot. The angle β_{Knee}, manipulated up-and-down motion of the robot's knee, is fixed at a constant value $\beta_{Knee} = 0\ (rad)$. At the end of the joint angle calculation phase, a set of joint angles $\theta = \{\ \alpha_{Hip}, \beta_{Hip}, \beta_{Knee}, \alpha_{RightShoulder}, \beta_{RightShoulder}, \alpha_{RightElbow}, \gamma_{RightElbow}, \alpha_{LeftShoulder}, \beta_{LeftShoulder}, \alpha_{LeftElbow}, \gamma_{LeftElbow}, \beta_{Head}, \gamma_{Head}, \gamma_{RightWrist}, \gamma_{LeftWrist}\ \}$ are released.

5.3.3 Boundary Constraint and Collision Check

Each of the robot's actuators has a limited range of rotation. To ensure that the calculated joint values satisfy the robot's physical configuration, joint angles θ are check with boundary constraint as given by Eq. 5.18. Here, θ_{i_min} and θ_{i_max} denote the lower and upper limits of the actuator θ_i. Finally, before releasing them to the Pepper robot, collision detection is conducted using the robot off-the-shelf model to prevent potential self-collisions.

$$\theta_i = \begin{cases} \theta_{i\_min}, & if \quad \theta_i \leq \theta_{i\_min} \\ \theta_i, & if \quad \theta_{i\_min} < \theta_i < \theta_{i\_max} \\ \theta_{i\_max}, & if \quad \theta_i \geq \theta_{i\_max} \end{cases} \tag{5.18}$$

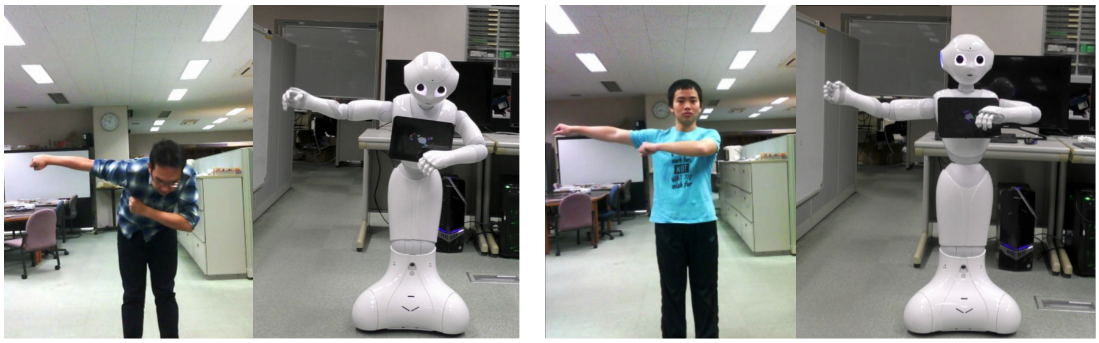
Chapter 6

Experiments and Discussion

In this chapter, we present a series of experiments conducted to evaluate the model of generating emotional gestures in chapter 3, the model for generating communicative gestures in chapter 4, and the transformation model illustrated in chapter 5. It is started by two experiment scenarios, described in section 6.1, in order to validate the transformation model. In section 6.2, an experiment was conducted to evaluate the integration between the model for generating emotional gestures and the transformation model. Finally, section 6.3 explains experiments conducted to evaluate the model for generating communicative gestures, and the transformation model.

6.1 Transferring Human Social Gestures into the Robot

In this experiment, the transformation model described in chapter 5, 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



(a) Human demonstration $H5$ (b) Robot imitated action $R5$ (c) Human demonstration $H6$ (d) Robot imitated action $R6$

Figure 6.1: The users stood in front of the Pepper robot and performed one-shot demonstration. The demonstrators' actions were imitated by the robot.

[45] or bodily expressions [12].

6.1.1 Experiment Scenario: Generating Robot Actions through One-shot Human Demonstration

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 [121], 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. 3.1. The encoded vector C captures the spatial-temporal information of motions as described in chapter 3. 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. 6.1. Hence, the closer the cosine distance to 1, the greater the similarity between the two vectors.

$$Similarity(C_a, C_b) = \frac{C_a \cdot C_b}{||C_a||\,||C_b||} \tag{6.1}$$

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.

Results and Discussion

The three demonstrators performed six actions combining the movements of their hip and arms, each of them demonstrated two actions. Table 6.1 shows the similarity between all pairs of demonstrator's actions calculated from Eq. 3.1 and Eq. 6.1. 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, as shown in Fig. 6.2, 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

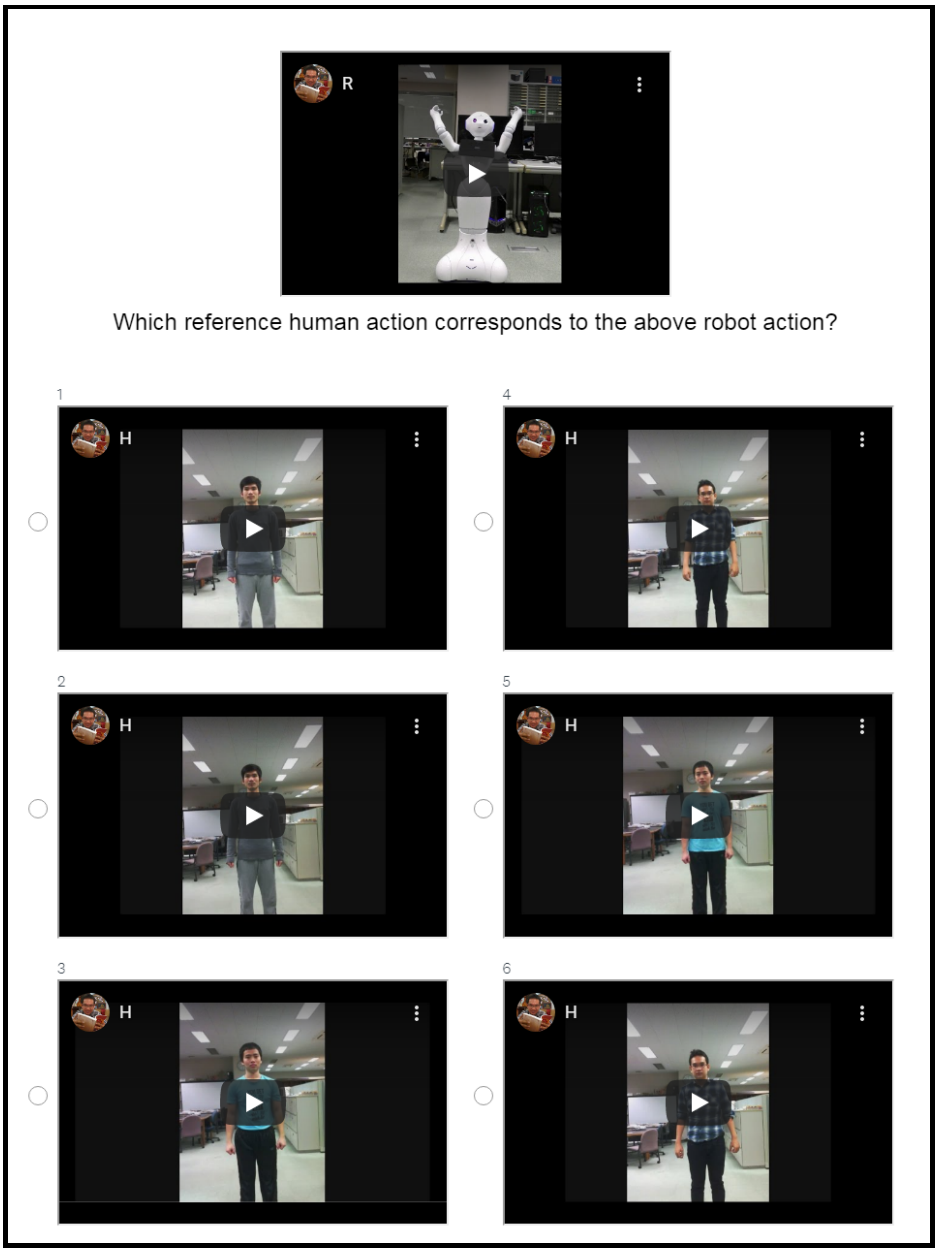


Figure 6.2: An experimental trial in the survey of the experiment. It is designed as a six alternative forced choice task where the observers select the most similar human action to the robot's action.

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

Table 6.1: Similarity between all pairs of human actions.

Action	*H1*	*H2*	*H3*	*H4*	*H5*	*H6*
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 6.2: Confusion matrix representing the recognition of six human actions (H) transformed into the robot model (R), normalized by the number of observers.

Action	*H1*	*H2*	*H3*	*H4*	*H5*	*H6*
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**

a visual angle of 11.17° (vertical) and 8.00° (horizontal). The viewing distance is approximately 70 *cm*. Table 6.2 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 6.1 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.

6.1.2 Experiment Scenario: Human Emotional Expressions Retained by Robot Motions

Experimental Setup

We conducted a study to evaluate whether the messages of human emotional gestures are retained by the robot motions using the UCLIC Affective Posture and Motion Database [66]. 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 6.1.1, 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 orders of trials were randomly presented to the observers. On each trial as presented in Fig. 6.3,

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. As shown in Fig. 6.4, 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.

Results and Discussion

Figs. 6.5a, 6.5c, 6.5e, and 6.5g show the key poses of the four selected human emotional gestures (*Happy, Sad, Fear, Angry*) chosen from the UCLIC dataset. Through the transformation model, those bodily expressions were converted to the Pepper robot motions considering the physical constraints as shown in Figs. (6.5b, 6.5d, 6.5f, 6.5h).

Subjective evaluations were conducted for both the human and robot emotional bodily expressions. Figs. 6.6a and 6.6b 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. 6.6a 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. 6.6b. Notably, the bodily expression *Angry* has the lowest recognition accuracy.

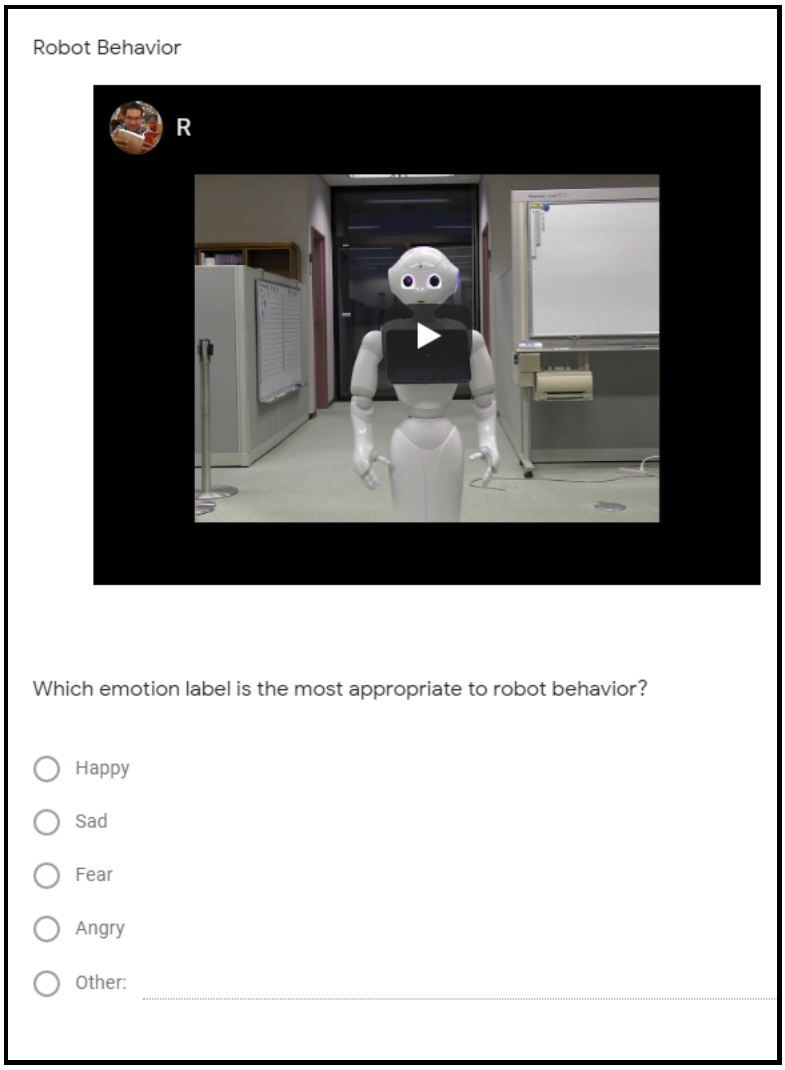


Figure 6.3: An experimental trial conducted in Experiment 6.1.2. The observers were asked to watch the robot's bodily expressions and choose the most appropriate emotion label from a list of five emotions - in a five alternative force choice task.

Fig. 6.5e 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 the 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

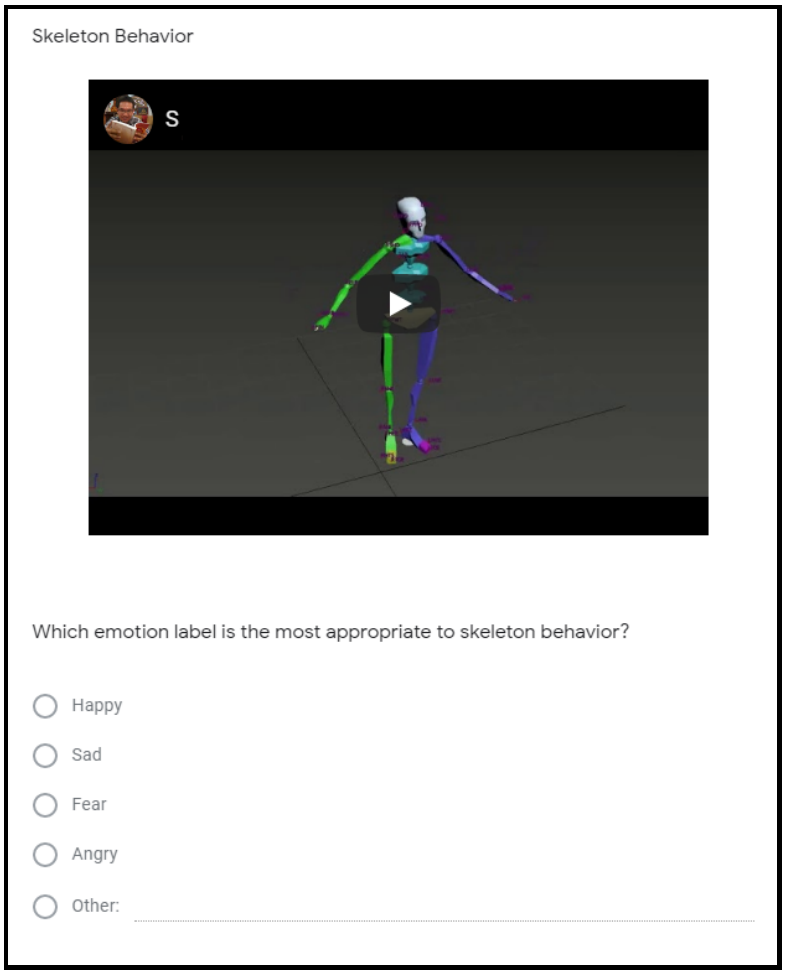
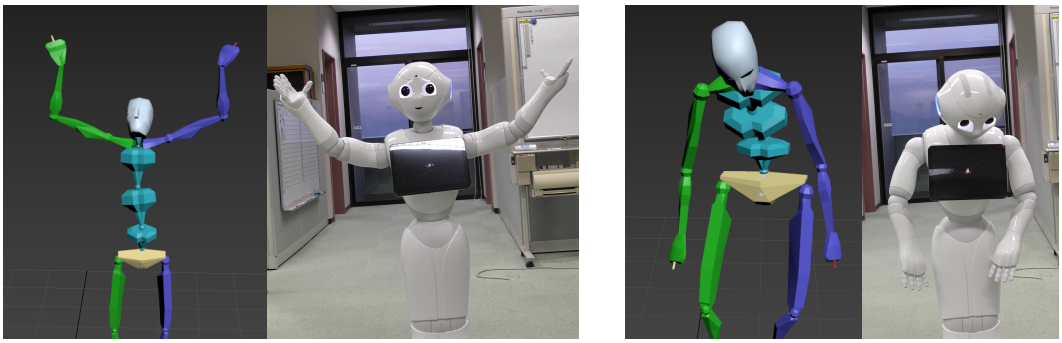
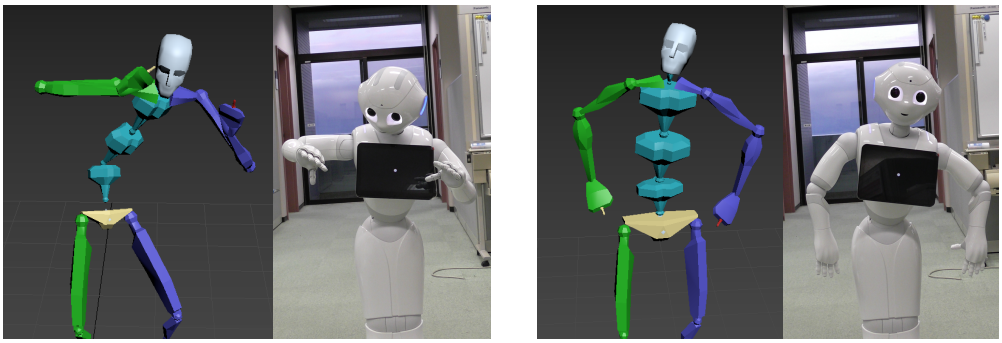


Figure 6.4: An experimental trial in the second part of survey of Experiment 6.1.2. It is designed as a five alternative force choice task.

joint β_{Knee} is set to a constant value of $\beta_{Knee} = 0\ rad$ (as the value at the initial position). Indeed, Fig. 6.7 indicates the absolute differences 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%



(a) human *Happy* (b) robot *Happy* (c) human *Sad* (d) robot *Sad*

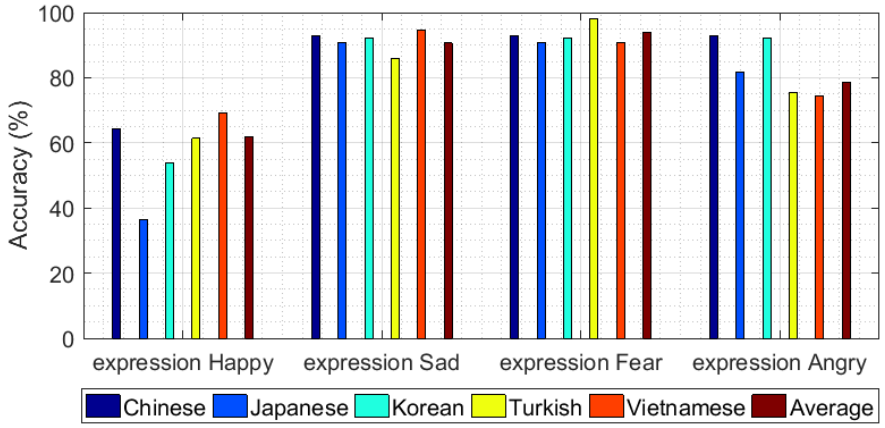


(e) human *Fear* (f) robot *Fear* (g) human *Angry* (h) robot *Angry*

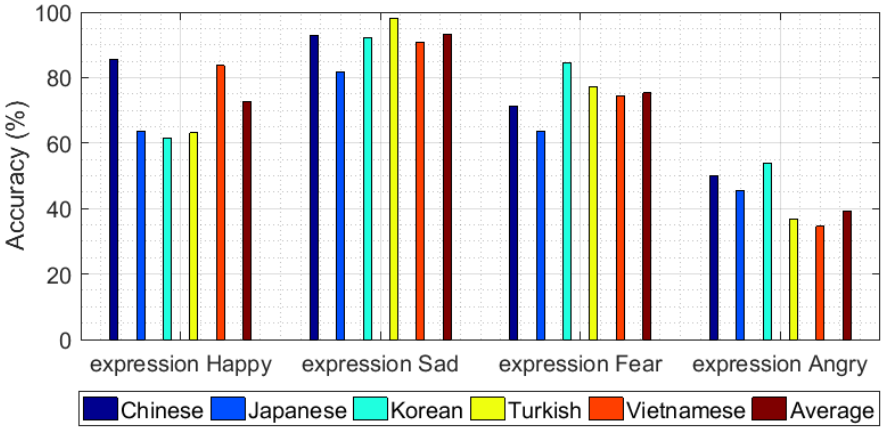
Figure 6.5: Selected human postures from UCLIC dataset visualized using Autodesk 3ds Max: Figs. 6.5a, 6.5c, 6.5e, and 6.5g represent the key poses of human bodily expressions. Figs. 6.5b, 6.5d, 6.5f, and 6.5h show the corresponding Pepper expressions.

for the human skeleton *Fear*.

As shown in Fig. 6.5a, the gesture *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



(a) Human bodily expressions



(b) Robot bodily expressions

Figure 6.6: 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.

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. 6.6b. On the other

Table 6.3: Recognition of emotional expressions of human skeleton normalized by the number of observers.

Emotional label	**Observers**				
	Happy	***Sad***	***Fear***	***Angry***	***Others***
Happy	**0.62**	0.03	0.12	0.14	0.09
Sad	0.01	**0.90**	0.01	0.01	0.06
Fear	0.01	0.01	**0.94**	0.01	0.03
Angry	0.05	0.04	0.01	**0.79**	0.10

Table 6.4: Recognition of emotional expressions of robot normalized by the number of observers.

Emotional label	**Observers**				
	Happy	***Sad***	***Fear***	***Angry***	***Others***
Happy	**0.73**	0.01	0.02	0.04	0.20
Sad	0.01	**0.93**	0.02	0.01	0.03
Fear	0.03	0.06	**0.75**	0.05	0.11
Angry	0.23	0.03	0.09	**0.39**	0.25

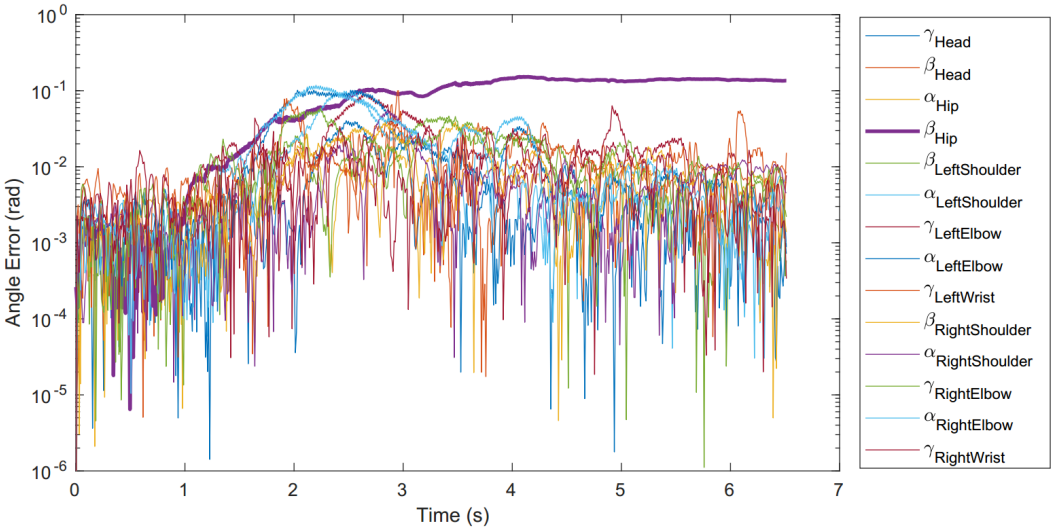


Figure 6.7: Absolute differences in joint angle values between the human expression *Fear* and the imitated one performed by the Pepper robot.

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 [122]. 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*.

6.1.3 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.

6.2 Incremental Learning to Develop Robot Emotional Gestures

In the previous experiments mentioned above, the efficiency of the transformation model to generate the robot's social cues through one-shot demonstration has been validated. The following experiment evaluates the integration of the model of generating emotional gestures illustrated in chapter 3 and the transformation model described in chapter 5. This integration is demonstrated through a scenario of three consecutive days of human-robot interaction. This scenario of interaction allows the target robot capable of learning human behaviors through long-term interaction and transforming them into its motion space, being the robot's emotional gestures.

6.2.1 Experimental Setup

The Scenario of Interaction

The experimental setup is given in Fig. 6.8, 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 model

[1] `http://doc.aldebaran.com/2-5/naoqi/peopleperception/alpeopleperception.html`
[2] `http://microsoft.com/cognitive-services/en-us/`

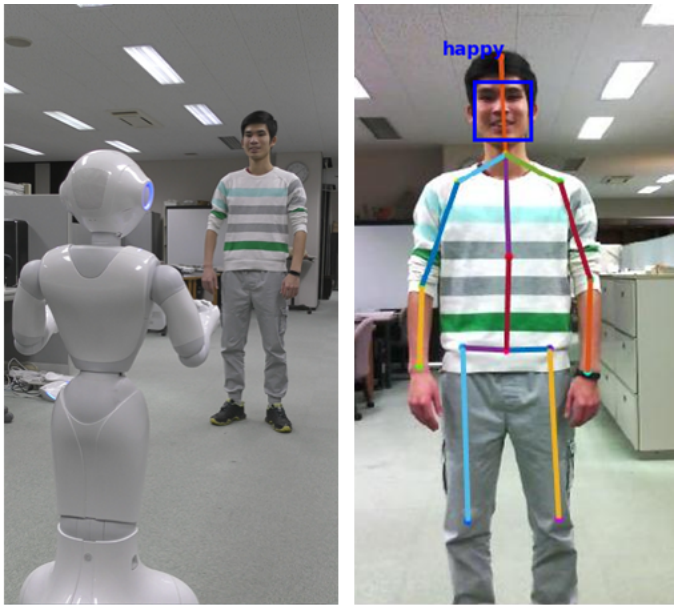


Figure 6.8: The scenario of Pepper's interaction for 3 consecutive days learning from the interacting partner's emotional behaviors.

as presented in chapter 3, 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.

Evaluation Criteria

This survey investigates the quality of the robot's emotional gestures 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

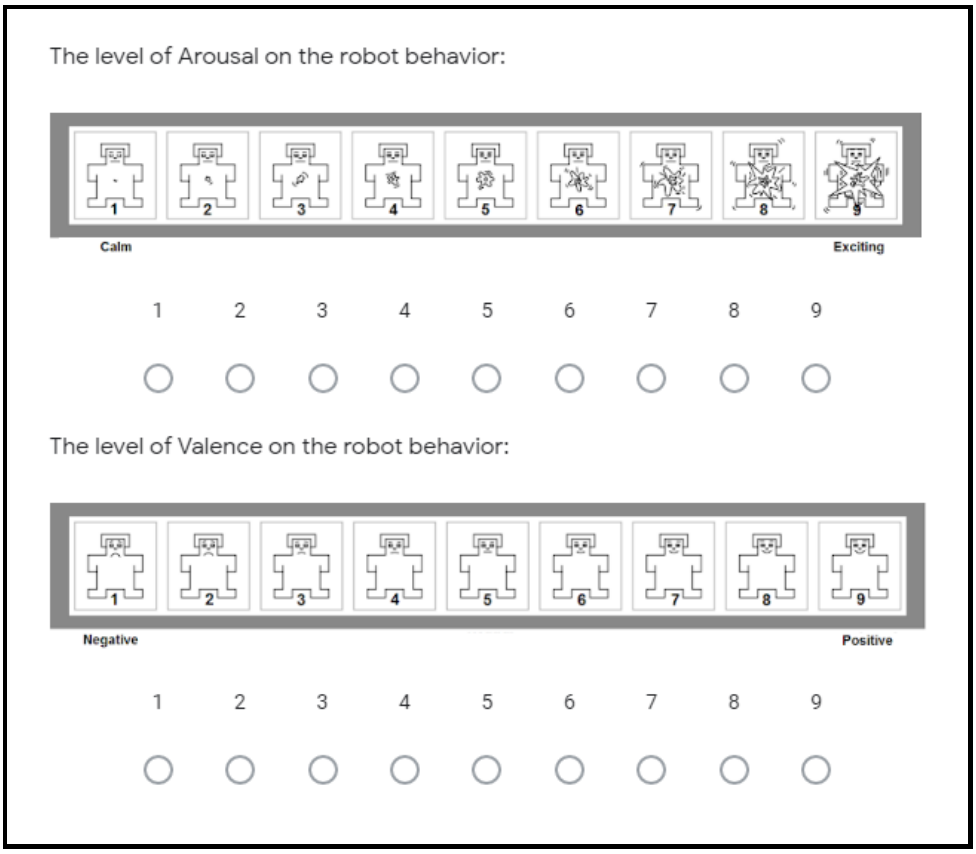


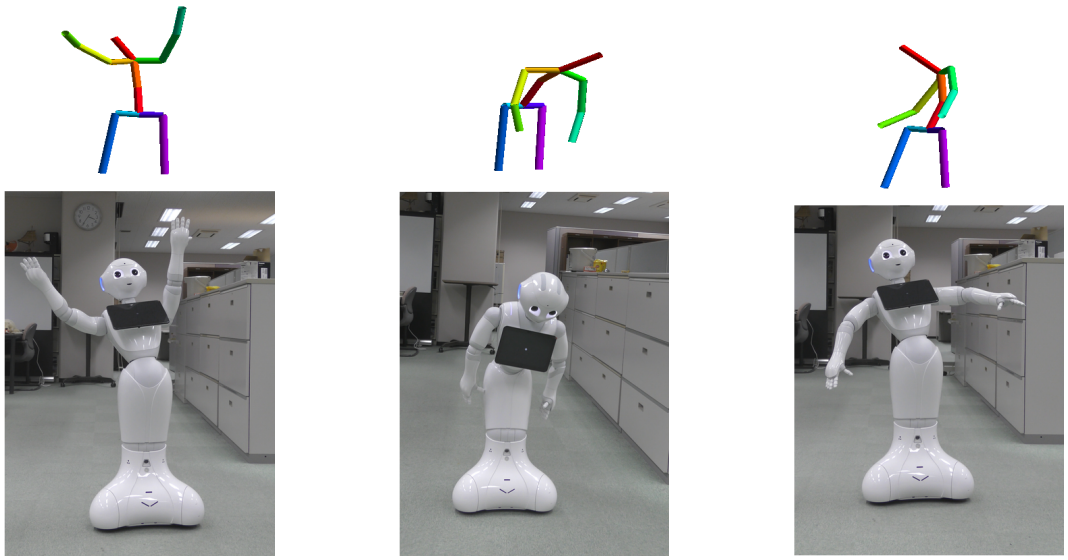
Figure 6.9: The observers rated appropriate Arousal and Valence values of the robot bodily expression using the Self-Assessment Manikin (SAM) nine-point scale

in English and most of them are not familiar with robots.

Firstly, the observers were asked to watch the robot's emotional gesture and choose the appropriate emotional label similar to the previous experimental setup mention in section 6.1.2. Then, the observers rated the appropriate value for Arousal and Valence using the Self-Assessment Manikin (SAM) nine-point scale [123]. Arousal and Valence are the dimensions on the Circumplex model of affect [124]. 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 [125, 126].

Table 6.5: 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



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

Figure 6.10: The key poses of Pepper emotional gestures produced using A_{rep} of the behavior selection phase.

6.2.2 Results and Discussion

Robot Bodily Expressions Generated Over Three Consecutive Days of Interaction

The model of generating emotional gestures 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 expres-

Table 6.6: The behavior selection phase on the third day. Using Eq. 3.15, 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

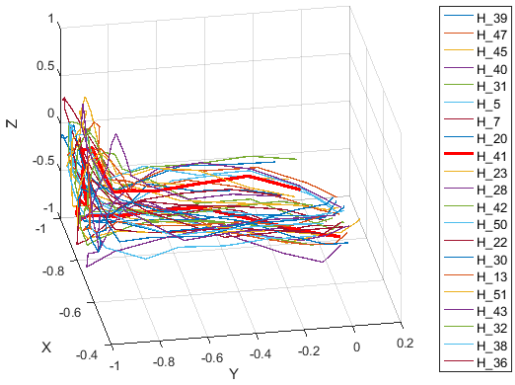
(b) $Cluster_i$ on emotion class *Sad*

	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

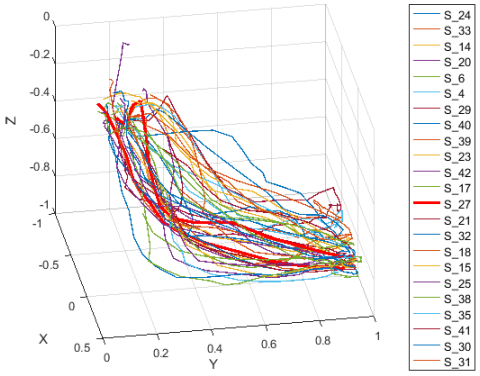
(c) $Cluster_i$ on emotion class *Fear*

	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

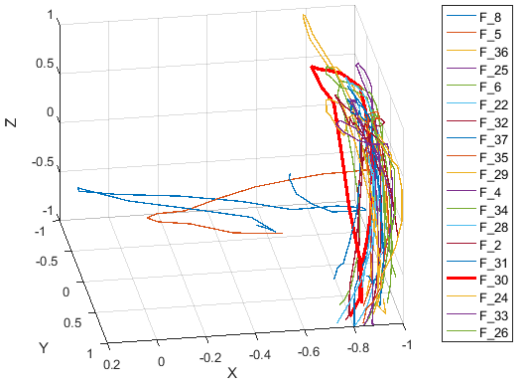
sions into the robot motions. This process was continuously repeated over three consecutive days as a part of the robot's social development. Fig. 6.12 shows the number of learned behaviors and the changes in the robot's emotional expressions over three days. More specifically, Table 6.6 represents the selected patterns from the behavior selection phase conducted on the last day. Based on the transformation model, the selected behaviors were converted to the robot motions, being the



(a) Emotion space *Happy*



(b) Emotion space *Sad*

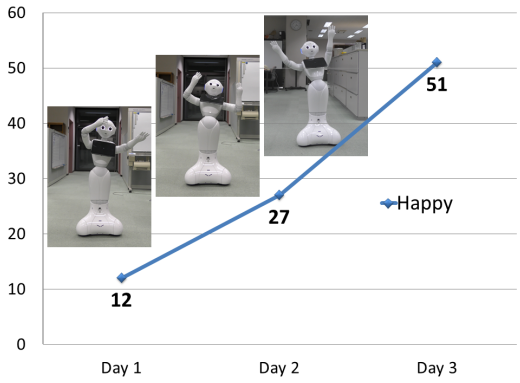


(c) Emotion space *Fear*

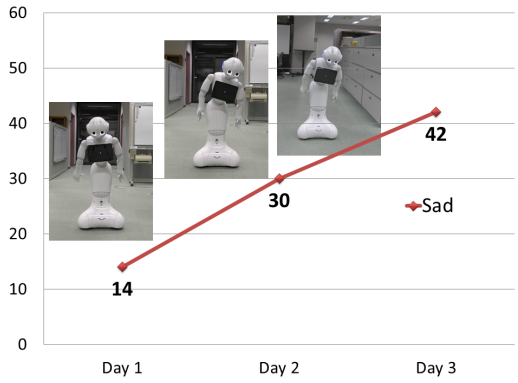
Figure 6.11: The trajectories of human left hand created by the patterns of Table 6.6. Eq. 3.15 selects the representative gesture A_{rep} the most consistent one in the cluster.

robot's emotional expressions. Fig. 6.10 shows the key poses of those behaviors.

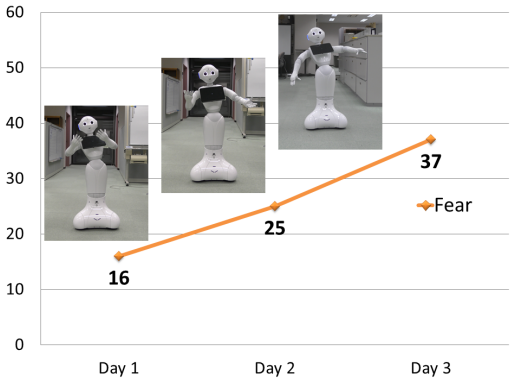
In our previous work [75], the training and clustering phase described in chapter 3 was evaluated with the Microsoft Research Cambridge-12 Kinect gesture dataset (MSRC-12) [72]. The experiment results as summarized in Table 6.5 indicated that SOM yielded better performance than DCS. Notably, the accuracy 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



(a) Emotion space *Happy*



(b) Emotion space *Sad*



(c) Emotion space *Fear*

Figure 6.12: Variational patterns of emotional behavior obtained through 3 consecutive days: Pepper robot incrementally learns and updates their emotional gestures day by day.

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. 3.15. Eq. 3.15 guarantees that the representative gesture A_{rep} is the most consistent one in that cluster. Tables 6.6a, 6.6b, 6.6c 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 6.6, the trajectories of the human left-hand are depicted in Fig. 6.11. Here, the movements of the hand were analyzed, since the hand movements are considered as the richest source of emotional body language [127]. Concerning the behavior selection phase as described in Table 6.6a, it is easy to notice that pattern H_41 satisfies Eq. 3.15. Visualizing the trajectories as shown in Fig. 6.11a, the trajectory created by the gesture H_41 is correctly located in the center of the cluster. As shown in Table 6.6c, 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. 6.11c 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. 6.10.

The Cultural Differences in the Perception of Robot Expressions

While the experiment results in section 6.1.2 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 emotional gestures on the last day as shown in Fig. 6.10 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 6.7 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

Table 6.7: 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.

Emotion 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 6.8: The recognition rate of robot expressions rated by 136 observers from 5 different cultural groups, normalized by the number of observers.

Emotion 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

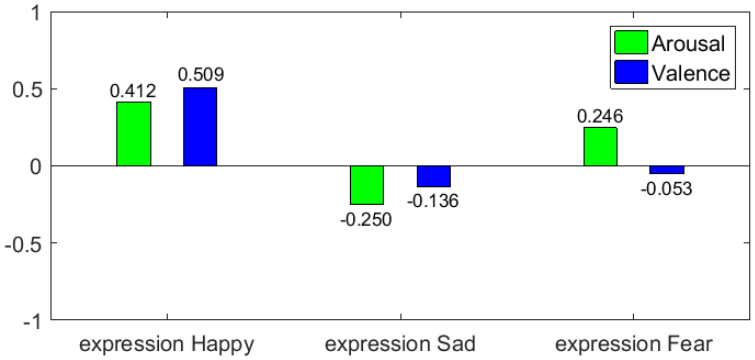


Figure 6.13: Mean values of Arousal and Valence rated by Vietnamese observers for robot expressions.

Arousal and Valence for the robot behaviors as shown in Fig. 6.13. Table 6.8 shows the recognition rate of 136 observers across five different cultures, while Figs. 6.14a and 6.14b represent the mean of Arousal and Valence assigned by the observers within individual cultural groups.

Table 6.7 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*.

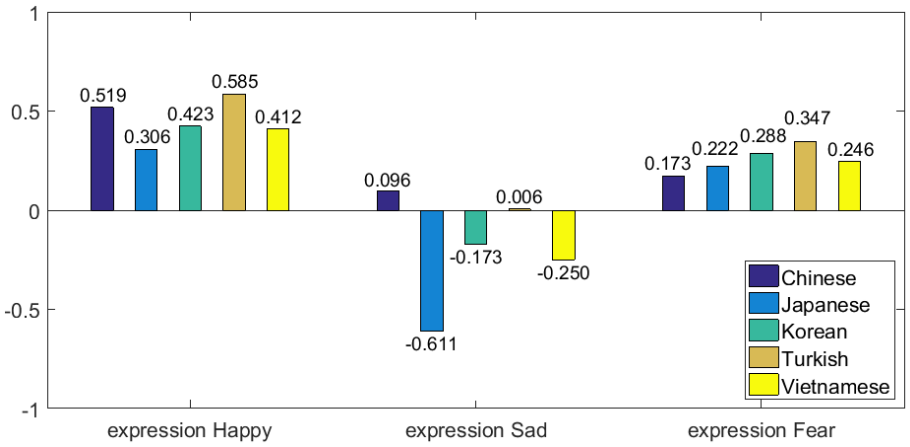
Table 6.9: The differences in Arousal and Valence for expressions *Happy*, *Sad*, *Fear* rated by Vietnamese observers. The third column indicates significantly different pairs.

Dimension	**ANOVA test**	**Post-hoc test**
Arousal	**significant diffs.** **p_value = 1.08E−14**	Sad-Happy = 4.02E-14 Sad-Fear = 5.27E-09
Valence	**Significant diffs.** **p_value = 3.8E−08**	Happy-Sad = 1.36E-07 Happy-Fear = 9.47E-06

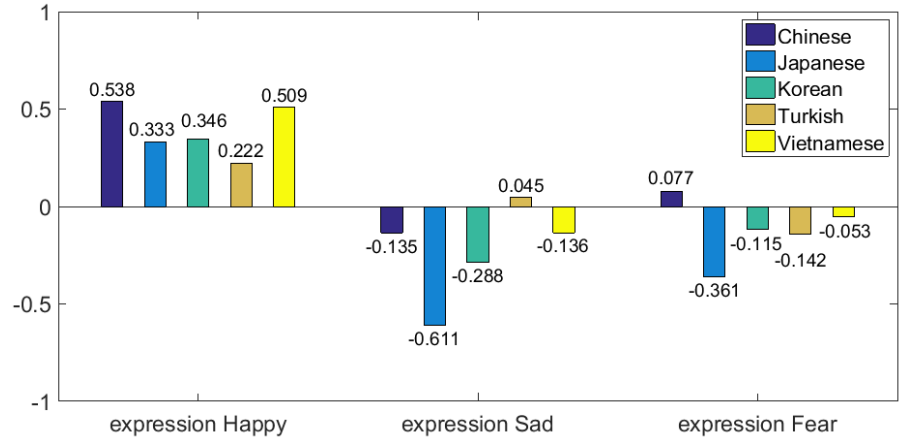
9% felt that the gesture might have another meaning such as *Excited*. The Pepper robot expressed the emotion *Sad* by slowly bending its upper body, keeping the hand positions lower than its Hip. 65% of observers assigned the label *Sad* to such Pepper motions. 19% rated it as another label like *Sorry*. 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 6.8 shows the recognition rate of 136 observers from five different cultures. In general, there were no significant differences noticed in the recognition rate of emotion labels 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 gestures 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



(a) Arousal dimension



(b) Valence dimension

Figure 6.14: Mean values of Arousal and Valence rated by people from 5 different cultures.

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. 6.13, 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

Table 6.10: 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.

differences. Table 6.9 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. 6.14a and 6.14b 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 6.10a and 6.10b 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 it. In general, the significant differences as mentioned above suggested that different cultural groups perceived the same emotional gestures of the robot in different ways.

6.2.3 Summary

In this experiment, the scenario of long-term human-robot interaction was conducted to validate the proposed learning frameworks. In the model of generating emotional gestures, 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's emotional gestures, 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.

6.3 Generating Communicative Gestures Synthesized with Robots' Speech

In this experiment, the model for generating communicative gestures illustrated in chapter 4 is validated on public datasets. It is followed by quantitative comparisons with the related works to verify the efficiency of our proposed framework to generate human actions synthesized with speech text. Finally, using the transformation model described in chapter 5, generated co-speech gestures are transformed into the target robot, being the robots' communicative gestures.

6.3.1 Experimental Setup

Dataset and Preprocessing

The designed framework for generating communicative gestures was firstly validated on the MSR-VTT dataset [128] as similar as conducted in [1]. This dataset consists of $2,822$ actions a_r and $31,863$ corresponding natural language descriptions d (one action could be associated with more than one description). As shown in Fig. 6.15, $a_r \in \mathbb{R}^{3\times8\times32}$ is a sequence of $T = 32$ skeleton frames representing the human upper body motion. Each frame S includes 8 joints defined in 3D Cartesian space. From the dataset, we filtered the actions whose joint positions are out of the range $[-1, 1]$. Concerning the Embedding Description, as mentioned in section 4.3.1, we used the encoder phase of the skip-thoughts model trained with the BookCorpus dataset [129]. As the BookCorpus dataset consists of $11,038$ books in a variety of topics, it allows the encoded vectors to effectively capture the semantics and syntax of the input sentences, without being biased toward any particular domain. Totally, $29,663$ pairs of actions a_r and corresponding descriptions d were obtained. For each a_r, we also collected the miss-matching description $\hat{d}$. The obtained data a_r, d, and $\hat{d}$ were split into 90% for training and 10% for testing.

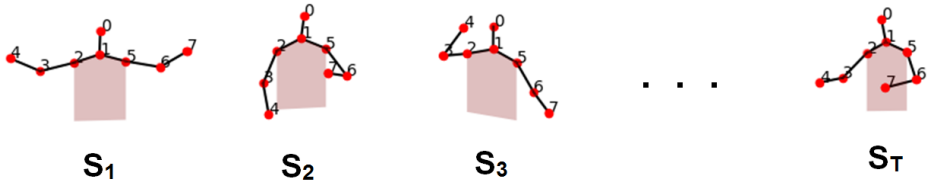


Figure 6.15: The action a_r consists of T skeleton frames in 3D Cartesian space. a_r is described by description d: "a person dances to a hip hop song".

Evaluation Metric

Consider that $a_r = [S_1, S_2, S_3, ..., S_T]$ is the real action associated with the description d, and $a_f = [S'_1, S'_2, S'_3, ..., S'_T]$ is the fake action synthesized with d. In order to verify the synthesis between a_f and d quantitatively, we used covariance description with temporal hierarchical construction [73] to evaluate how similar the generated action a_f and the real action a_r are. Given a_r and a_f as the inputs, Eq. 6.2 encodes them as the corresponding feature vectors C_r and C_f, respectively. Here, $\overline{S}$ is the sample mean of S_i computed over the time T and $^\intercal$ represents the transpose operator. This feature vector efficiently captures spatio-temporal information of action over the time sequence, it has been used for action recognition tasks [73] and unsupervised learning tasks [130]. Finally, the similarity between C_r and C_f is measured by cosine similarity as given in Eq. 6.3.

$$C = \frac{1}{T-1}\sum_{i=1}^{T}(S_i - \overline{S})(S_i - \overline{S})^\intercal \tag{6.2}$$

$$Similarity(C_r, C_f) = \frac{C_r \cdot C_f}{||C_r||\,||C_f||} \tag{6.3}$$

6.3.2 Results and Discussion

Variety of Actions Conveying a Certain Context Input

From the training data, the real action a_r, the matching description d, and the miss-matching one d were fed to the designed network with the batch size 100. The dimension of the noise vector z is 100. The Adam optimizer [131] with the momentum 0.5, and the learning rate 2×10^{-5} was applied for both G and D network.

Figure 6.16: Skeleton sequence of generated action for "a young woman demonstrates example of lifting exercises."

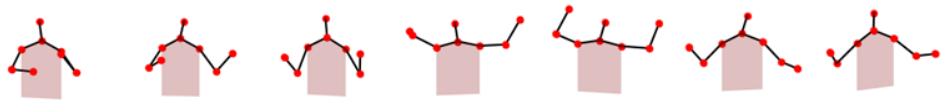


Figure 6.17: Generated action for "a girl practices lifting exercise at the gym."

Figure 6.18: Generated action for "a woman performs weight lifting exercises."

Figure 6.19: Generated action for "I was practicing lifting exercises at the gym."

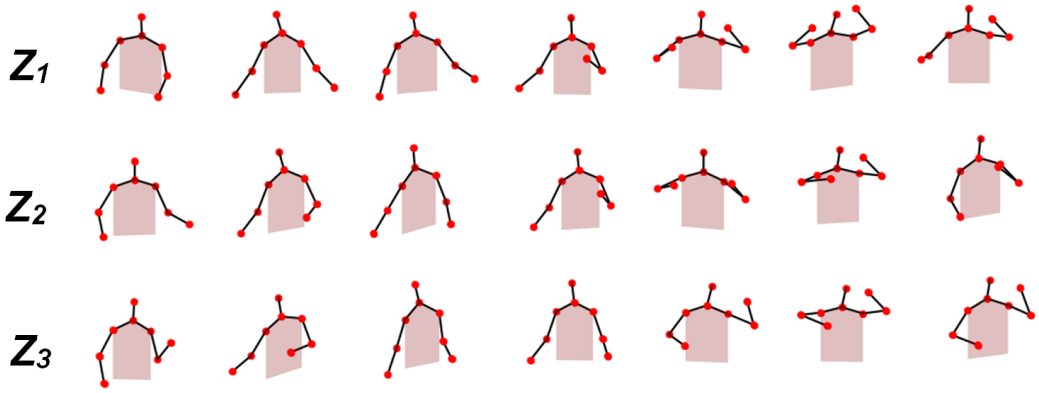


Figure 6.20: Generated actions for"one girl is dancing to music". Those are produced from the noise vector z_1, z_2 and z_3, respectively.

The Discriminator and Generator were sequentially trained for 700 epochs.

Once the training process is completed, at the testing phase, we firstly fed different annotation texts d and a fixed noise vector z to the Generator network.

The matrix x_f produced by G was decoded to a_f. The generation action a_f is defined in 3D Cartesian space, this bodily expression aims to convey the meaning of context d. Fig. 6.16 illustrates the generated action of the proposed model by feeding an input "a young woman demonstrates example of lifting exercises", which is included in the testing data. The action looks like a person is lifting two arms over the shoulder two times. Moreover, we also tested the two modified versions of that sentence such as "a girl practices lifting exercise at the gym" and "a woman performs weight lifting exercise". The resulting actions are presented in Fig. 6.17 and Fig. 6.18, respectively. It is noticed that those actions look like a person is lifting something by pushing their hands up over the shoulder several times. A closer look at Fig. 6.16, 6.17, and 6.18 show that skeleton frames of those actions are not exactly matched to each other at a certain timestamp. However, generated bodily expressions seem to be similar over the time sequence. In a second demonstration, the same text description d "one girl is dancing to music" was given to the Generator network with different noise vectors z_1, z_2, and z_3. The generated actions are displayed in Fig. 6.20. It can be seen that those actions are demonstrated by a similar bodily expression over the whole time sequence. However, the amplitudes of those motions are slightly different at a certain timestamp. Overall, the results demonstrated above suggests that the G network does not merely memorize and reproduce the data. It is able to generate a diverse set of actions to convey a particular meaning of context input. Taking into account scenarios of human-robot interaction, this capability would allow social robots to perform novel behaviors over time, which positively contributes to the user's engagement during interaction [19].

Quantitative Evaluation of Generated Actions

The real action a_r is correctly synthesized with the text description d. Thus, it is reasonable for evaluating actions produced by the G network by measuring the similarity between a_r and a_f, since those are synthesized with the same description d. Notice that a_r and a_f could express the same meaning over the time sequence, although their corresponding skeleton frames are not exactly matched each other at a certain timestamp. The evaluation metric suggested in 6.3.1 satisfies such

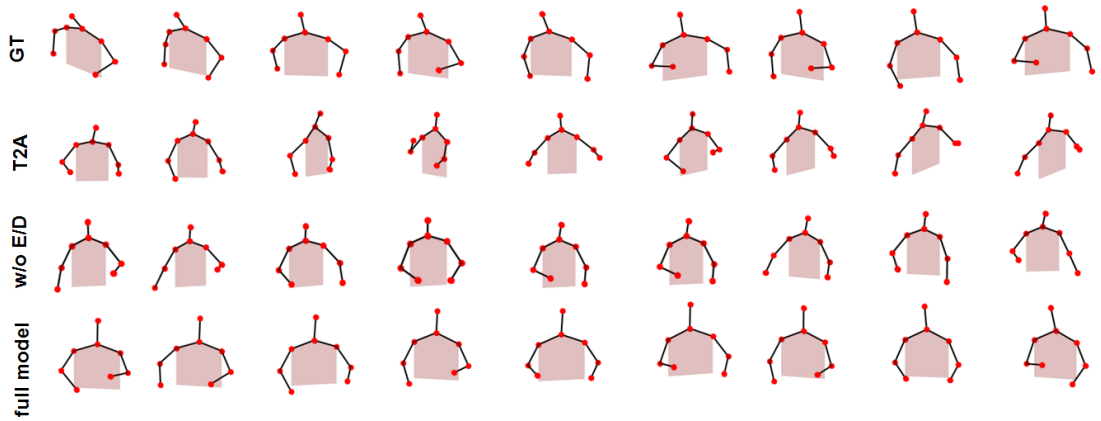


Figure 6.21: Comparison with the real action (GT) for "a sprinter is sprinting on the track with his head down": Text2Action (T2A) [1], the model without Action Encoder/Decoder (w/o E/D) [2], and the fully implemented model (full model) [3].

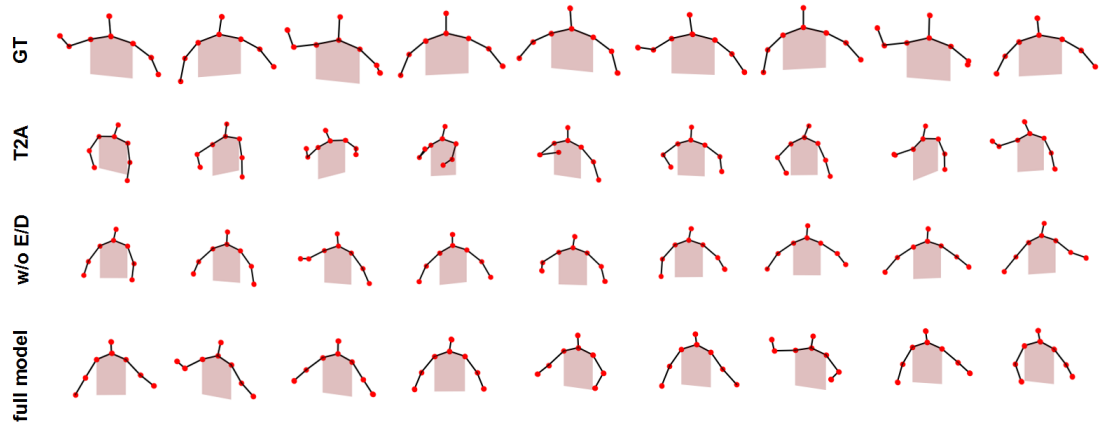


Figure 6.22: Comparison with the real action (GT) for 'a man skiing up a hill at a competition".

requirement. Here, we sequentially fed text descriptions of the testing data to a given G network. Both the generated and real actions were plugged into Eq. 6.2 and Eq. 6.3 for measuring their similarity.

To quantitatively verify the differences between our proposed network and the related approach - Text2Action [1], we trained their proposed network again on this training data while keeping the same training parameters as suggested by the authors. Additionally, we also verified the efficiency of the action generation frame-

Table 6.11: Similarity comparison among Text2Action [1] (T2A), the model without Encoder/Decoder [2] (w/o E/D), and the fully implemented model [3] (full model)

	Text2Action	w/o E/D	full model
Average similarity	0.4196	0.5060	0.5287

work without the Action Encoder and Decoder, which is described in our previous work [2]. Specifically, the raw action a_r was fed to the designed network without encoding. Then, the action a_f is generated from G without passing through the Action Decoder. Table 6.11 presents the similarity between the real actions of testing data and actions generated from Text2Action, our model without Action Encoder/Decoder, and fully implemented model, respectively.

Table 6.11 indicates that by feeding the same text descriptions, the generated actions produced by our networks are more similar to the real ones. Thus, our generated data are more connected to the input sentences. It should be emphasized that our D network is trained to differentiate between data generated by G and the real training data taking into account the description d as similar as applied in [1, 2]. Additionally, D is trained to detect the error when the real action is associated with the miss-matching text $\hat{d}$. This strategy enables the Discriminator capable of evaluating the synthesis between a given action and a conditional input in a more efficient way. On the other hand, Table 6.11 indicates that the fully implemented framework yields higher accuracy than the one without Action Encoder and Decoder. The experiment showed that by feeding the raw input a_r to the framework as applied in [2], the training process was faster since the Action Encoder encodes a_r as x_r, which is the higher dimension matrix. However, by distributing the relative joints near each other as in x_r, it allows the spatial and temporal information of a_r to be represented better. Thus, D could detect the motion properties of the input action faster and more efficiently. Consequently, D provides more informative feedback to G, for optimizing the generated action. Fig. 6.21 displays an example of feeding a sentence "a sprinter is sprinting on the track with his head down" to the three G networks. The real sample indicates a person that is pumping two hands up and down while the head is bent down slightly. Although the posture of bending his/her head down is unsuccessfully ex-

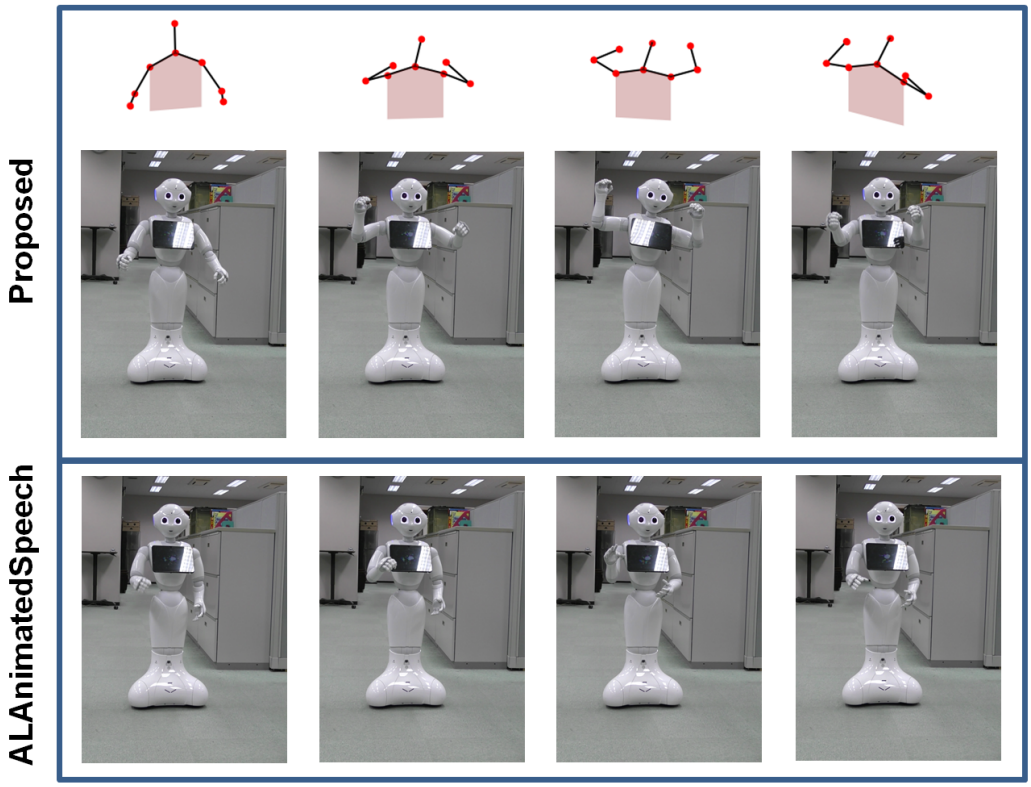


Figure 6.23: Differences on gestures between the proposed approach and ALAnimatedSpeech for describing input "one girl is dancing to music"

pressed neither by the three generated actions, those actions look like persons are pumping two hands up and down several times. Especially with the fully implemented model, the action is more natural and similar to the real one. A similar finding can be seen on the generated actions to convey "a man skiing up a hill at a competition" shown in Fig. 6.22. The ground truth action demonstrates a person is pumping their right hand several times and leaning forward. His or her two hands are spreading out for maintaining balance. The result displayed in Fig. 6.22 indicates that such motion features are better presented on our generated actions, especially the one produced by our fully implemented model.

Transforming Generated Actions into the Pepper robot

The generator G produces the action a_f defined in 3D Cartesian space. Through the designed transformation model, which is described in chapter 5, the gen-

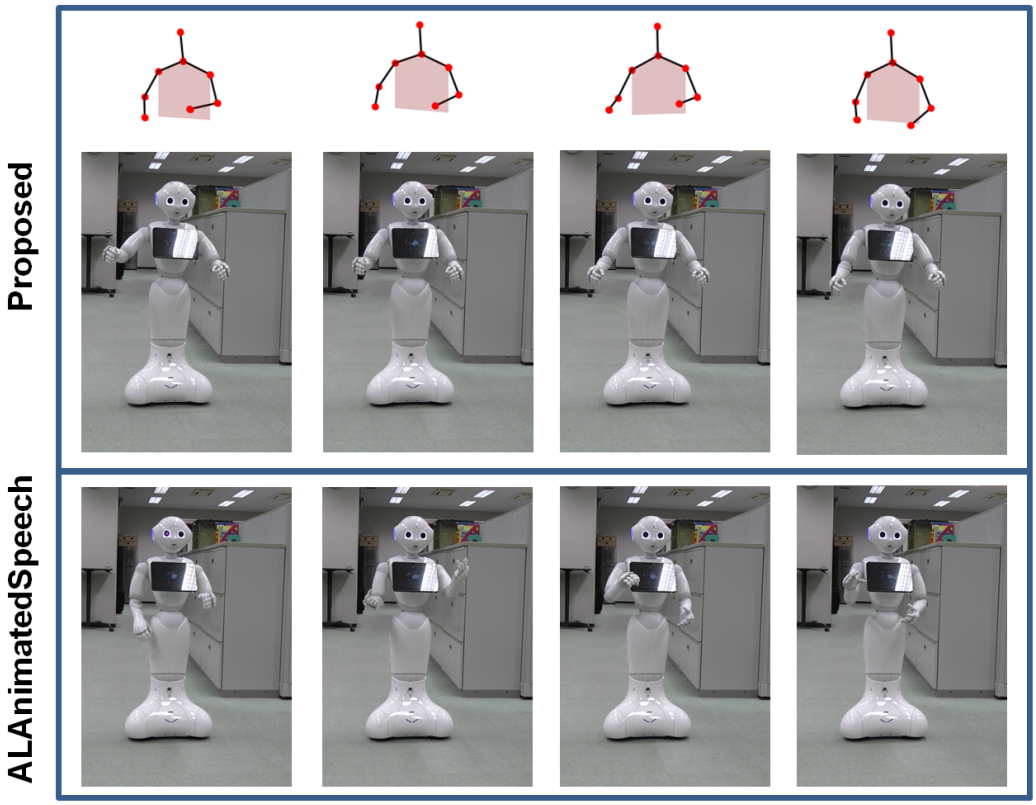


Figure 6.24: Generated actions for "a man is driving his motorbike on the street ".

erated action is converted to a set of corresponding joint angles, including $\theta = \{\alpha_{RightShoulder}, \beta_{RightShoulder}, \alpha_{RightElbow}, \gamma_{RightElbow}, \alpha_{LeftShoulder}, \beta_{LeftShoulder}, \alpha_{LeftElbow}, \gamma_{LeftElbow}\}$, for controlling the upper bodily expression of the target robot. The robot's physical constraints were taken into account during this transformation process. Notice that we used the robot's on-board module ALTextToSpeech [3] to enable robot to utter the input sentence d while performing the action a_f. Concerning the robot's off-the-shelf module, the robot's NAOqi API ALAnimatedSpeech[4] is provided to endows the Pepper robot talk in an expressive way. As the result, in order to qualitatively discuss the differences between our proposed approach and the ALAnimatedSpeech, the same speech text d was feed into the robot's module. Fig. 6.23, 6.24, 6.25, and 6.26 present the robot's ges-

[3] http://doc.aldebaran.com/2-5/naoqi/audio/altexttospeech.html
[4] http://doc.aldebaran.com/2-5/naoqi/audio/alanimatedspeech-api.html

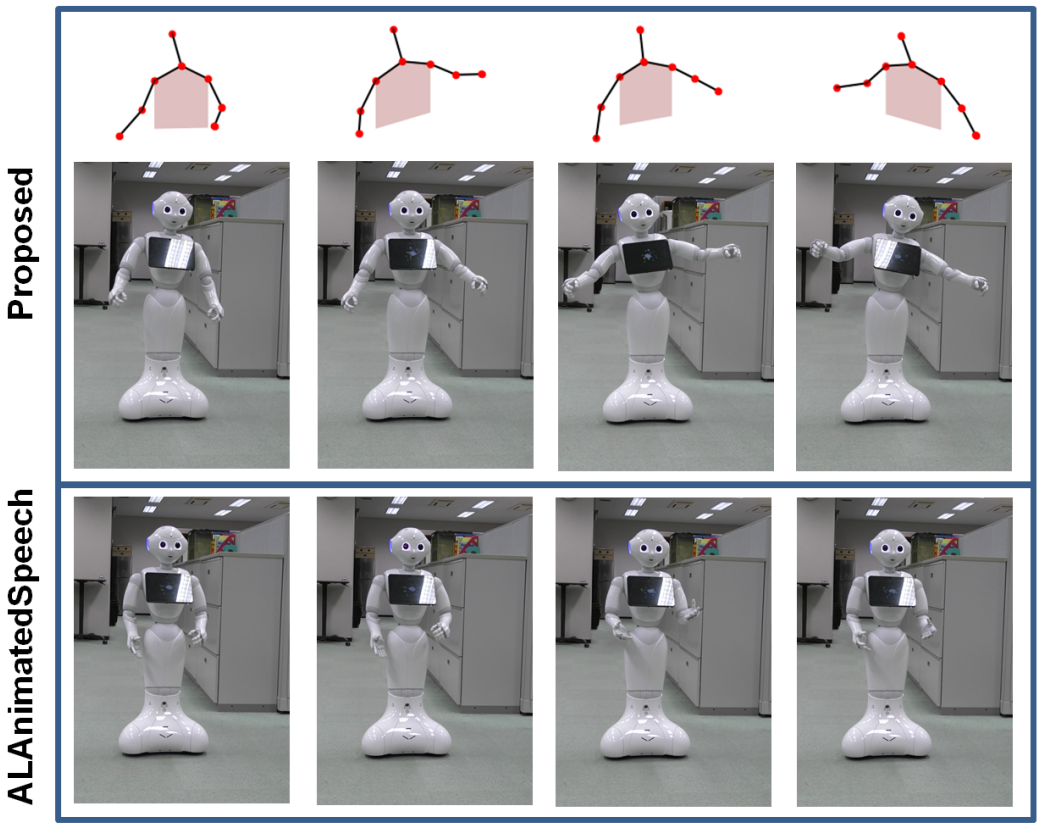


Figure 6.25: Generated gestures for "a man rides on the surf board in the water".

tures produced by our approach and the robot's off-the-shelf module. As shown in Fig. 6.23, our generated action could be observed as a person expresses something by performing energetic movements of their two hands. The same message of bodily expression could be observed in the robot's action produced by our proposed approach. Alternatively, it is expressed by slight movements of the robot's hands when feeding an input "one girl is dancing to music" into the ALAnimatedSpeech module. In Fig. 6.24, our generated action is demonstrated by a person constantly hold something in front of their body. On the other hand, it is difficult to interpret the meaning of bodily expressions generated by the robot's on-board module. In general, our experiments noticed that most of the Pepper robot's actions generated by ALAnimatedSpeech are not appropriately fit to the spoken texts, it is always the case that generated actions are expressed by slight movements of the robot's hands. Somehow, those gestures could be understood as a person is describing or

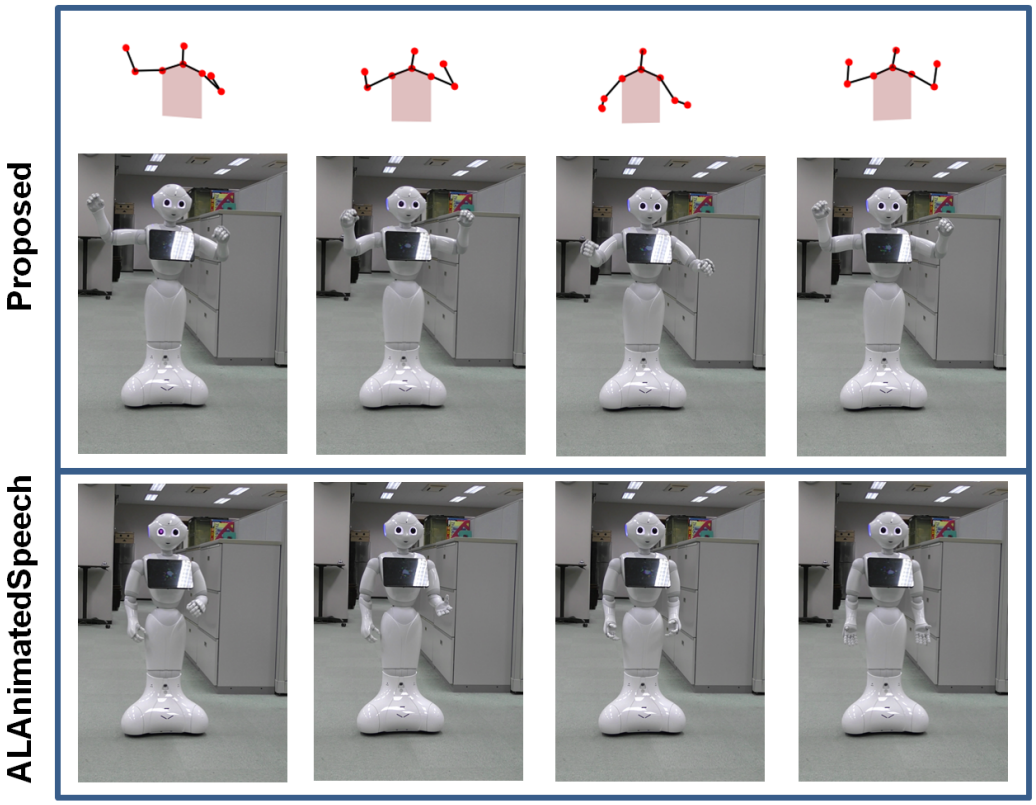
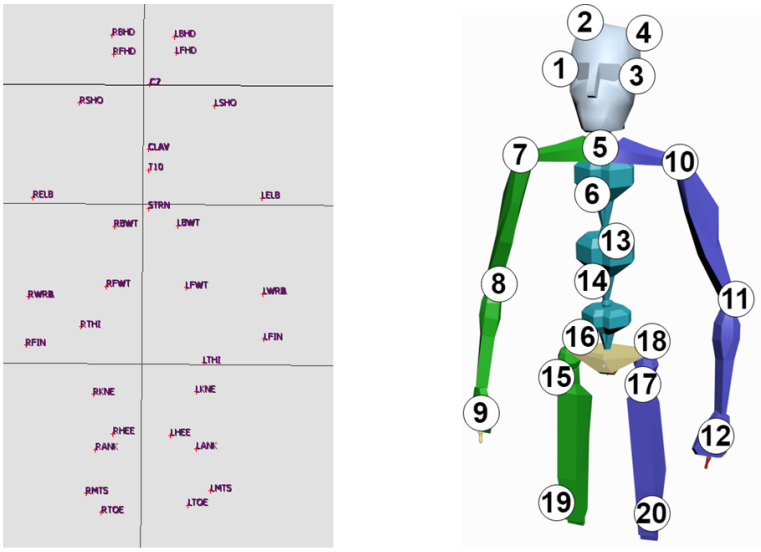


Figure 6.26: Generated gestures for"a young woman demonstrates example of lifting exercises".

presenting something as displayed in Fig. 6.25 and 6.26. It is important to notice that ALAnimatedSpeech consists of a set of actions handcrafted by animation experts to ensure the human-likeness and familiarity of the robot's gestures to human perception. By injecting an input text to ALAnimatedSpeech, a random action could be produced if certain keywords are not detected from the input. As the result, taking into account the use of bodily expressions for emphasizing the verbal content of the robot's speech, it suggests that the robot's on-board module can only produce stereotypical behaviors in a limited number of contexts.

6.3.3 Demonstration on a High Dimensional Dataset



(a) Raw optical marker data (b) Skeleton visualization

Figure 6.27: The left figure shows the raw motion capture data of the KIT dataset. We collected 20 markers capturing the motion of upper body and knees, they are visualized as human skeleton model as shown in the figure on the right side.

Preprocessing

In this experiment, we used the Karlsruhe Institute of Technology (KIT) whole-body motion dataset [132], and the corresponding natural language annotations [133]. The KIT motion dataset provides a rich corpus of human whole-body motion in a wide range of motion types. The selected data contains $2,127$ motions captured by 53 optical markers in 3D at the frequency of 100 Hz. Since this research focuses on generating the motions for the humanoid robot Pepper, only 20 markers capturing the motion of the human upper body and knees were selected out as the raw data as illustrated in Fig. 6.27. Noticed that the knees were included in order to compute the robot's hip joint angles. Each selected action $a = [S_1, S_2, S_3, ..., S_T]$ consists of a sequence of skeleton frames over a period of time T. At the frame i ($i \leq T$), $S_i = [x_1, x_2, .., x_{20}, y_1, y_2, .., y_{20}, z_1, z_2, .., z_{20}]$ ($S_i \in \mathbb{R}^{60}$) is the 60 dimensional vector that defines the positions of 20 joints in Cartesian space. Fig. 6.27b

shows the visualization of 20 selected motion capture data. The Autodesk 3ds Max is used to visualize human motion in a skeleton model. On the other hand, spelling errors in natural language annotations describing the demonstrative actions were corrected. With the $5,136$ usable annotation samples from the dataset (one action could be associated with more than one annotation), each description d was associated with the corresponding motion a. Similar to the previous experiment setup, we used the encoder phase of the skip-thoughts model, which was trained with the BookCorpus dataset [129], for generating the embedding description. In terms of the demonstrative actions, as they were recorded by the optical-based motion capture systems, the positions of markers highly depend on the camera coordinates. Thus, the joint positions were constructed with respect to the top-chest coordinates. On the other hand, the sizes of demonstrators are different from the training samples. Therefore, the actions were normalized to have the variance 1. Afterward, the motions were downrated to 10 Hz and padded to have an equal length of 240 frames. Totally, $51,360$ pairs of motions and descriptions were obtained. We split it into 90% for training and 10% for testing.

Identification of Human Joint Spatial Configuration

The motions and the corresponding natural language annotations from the training set were fed into the designed model with the batch size 100. The Adam optimizer [131] was used at the learning rate 2×10^{-5} for both the Generator and Discriminator network. The model was trained until Epoch $1,200$. During the first 30 epochs, only the Discriminator was trained. After that, both D and G were sequentially trained. In order to monitor intermediate motions of x, the same description d and noise z were given to G during the training process.

Different to the co-speech action $a_f \in \mathbb{R}^{3\times8\times32}$ trained with MSR-VTT dataset [128] that illustrated in the previous experiment. Here, the generated action $a_f \in \mathbb{R}^{3\times20\times240}$ consists of 240 skeleton frames, each frame contains 20 joints defined in 3D Cartesian space. It is revealed that the designed framework is able to cope with such high "resolution" actions, as illustrated in Fig. 6.28. At the beginning of the training phase, G could not capture the spatial configuration of the training samples. Because of that, the generated gestures at Epoch 10 and Epoch 20 do not

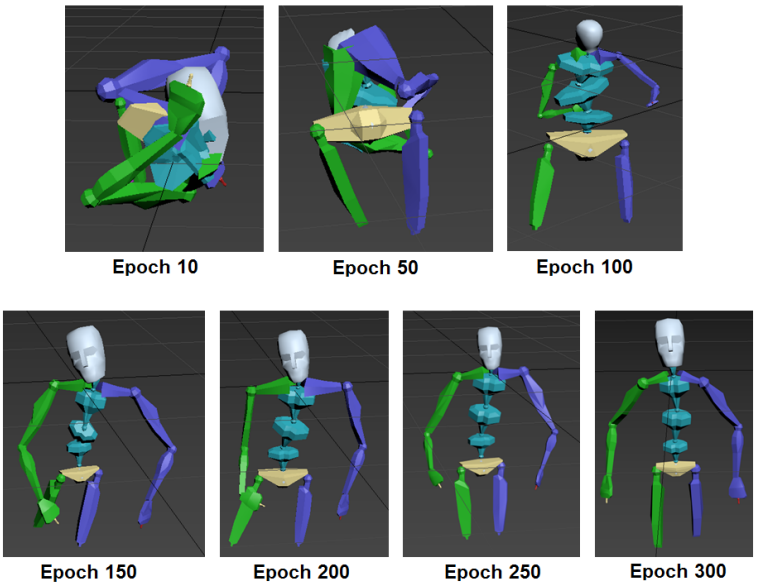
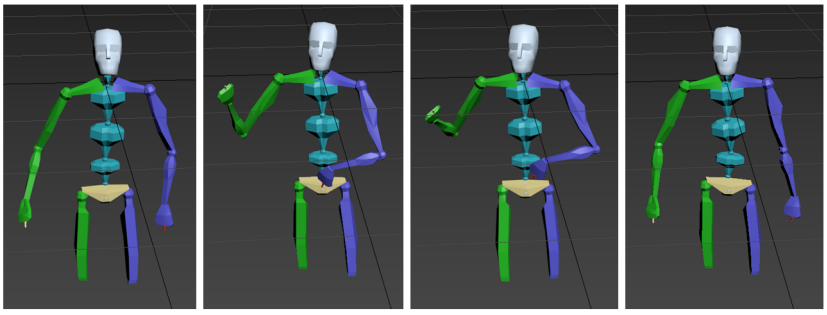


Figure 6.28: Throughout the training process, the Generator model imitated the human joints distribution so the generated poses looked more human-like.

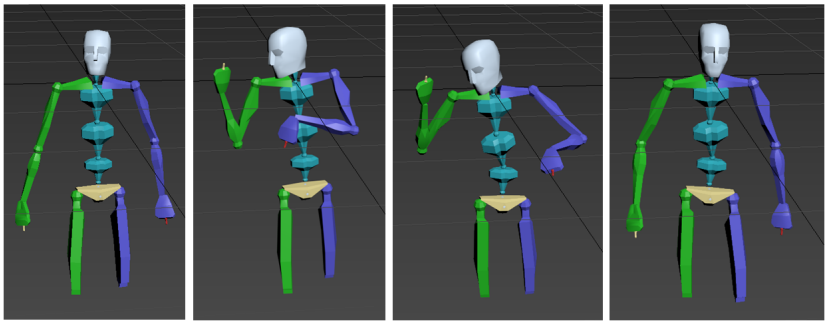
look like the shape of the human body. Starting from Epoch 100, G ameliorated the human joint configuration coordination problem and produced more natural human-like poses. At Epoch 300, the generated pose was well-proportioned as seen in Fig. 6.28. Hence, throughout the training process, the Generator was able to learn the coordination of human joint configurations. By the end of the training phase, G could generate the human body properly and symmetrically.

Generated Gesture Synthesized with Input Descriptions

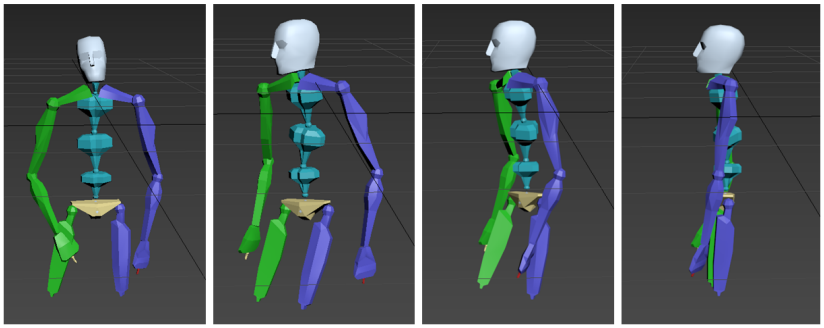
Figs. 6.29a, 6.29b, 6.29c, and 6.29d show the generated gestures produced by G network be feeding corresponding text descriptions as the input. Those sentences are included in the testing data. It is clear that by generating co-speech actions defined in a higher dimension, sophisticated contexts of input sentences could be expressed transparently. A closer look at the action "A human is playing a guitar", and "A human is playing violin" shown in Fig. 6.29a and Fig. 6.29b, it is suggested that without the head movements for holding a violin as displayed in Fig. 6.29b, to some degree, such two gestures would be similar to each other. Indeed, by equipping the hip and knee joints on the action output, locomotion actions as in



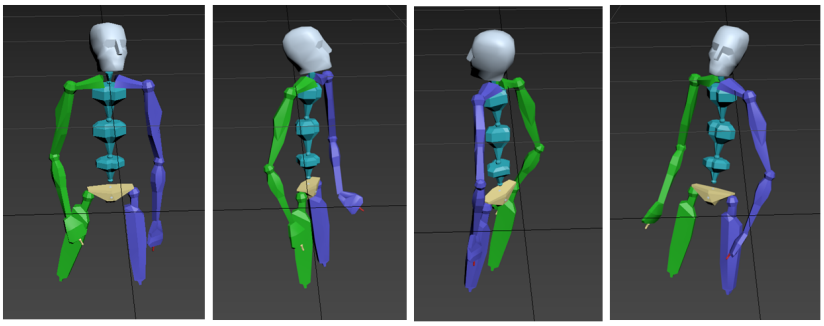
(a) “a human is playing a guitar”



(b) “a human is playing violin”

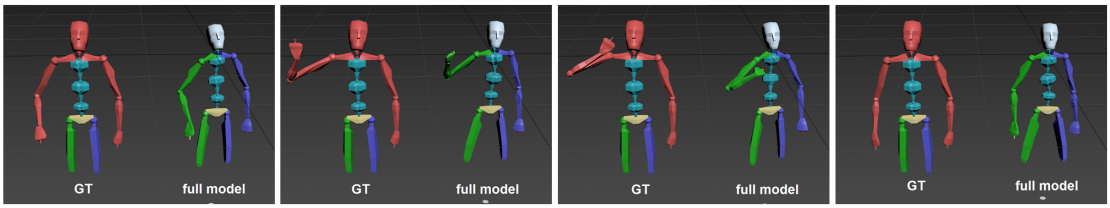


(c) “a person walks and turns to the right”

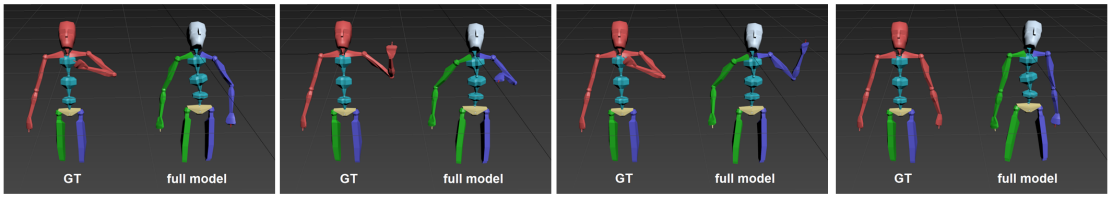


(d) “a human walks in a circle counterclockwise“

Figure 6.29: Generated human-like gestures synthesized with input sentences.



(a) “a person waves with its right hand“



(b) “a person waves with the left hand“

Figure 6.30: Comparison between the ground truth actions (GT) and the generate ones produced by our fully implemented model (full model).

Fig. 6.29c, and 6.29d could be expressed in a transparent manner. Noticed that at the preprocessing step, the human joint positions were constructed reference to the top-chest coordinates. This configuration makes the Generator G always tries to keep the position of the top-chest at the same position over the time sequence. As the results, it could be observed that actions presented in Fig. 6.29c, and 6.29d, look like a person is turning around while the position of their top-chest remained unchanged. Fig. 6.31 shows the 2-dimensional tSNE projection of a_f, each plot presents a generated motion. Based on the given description d, a_f could be categorized into several different motion types. In general, it can be seen that generated actions belong to a same motion type are grouped into a same cluster. However, the separation of locomotion actions (e.g. walking, running, etc.) is less clear. It is suggested that lower-body movements are ignored in the generated actions a_f, thus, bodily expressions of such motions are less accurate compared to the generated upper body motions (e.g waving, dancing, etc.).

Fig. 6.30 display two actions generated from our fully implemented model [3] as described in section 6.3, and the ground truth actions (GT). As presented in Fig. 6.30a, with the input text *”A person waves with its right hand”*, the real action and the generated one from our model are similar to each other. On each of the two actions, the first frame shows a human pose at the upright position,

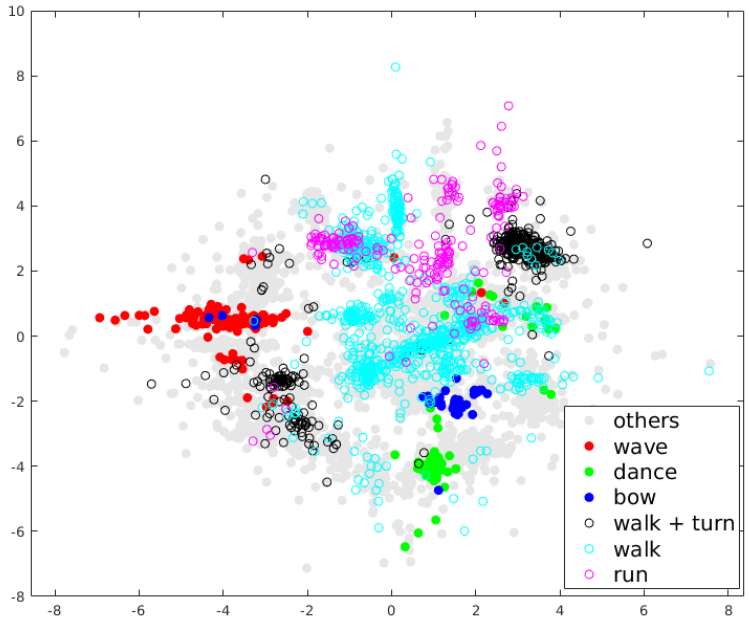
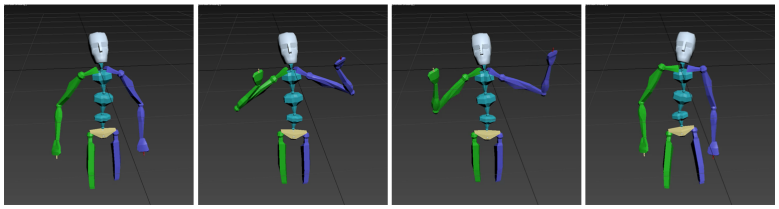


Figure 6.31: 2-dimensional tSNE projection of generated action a_f, colored by their motion types.

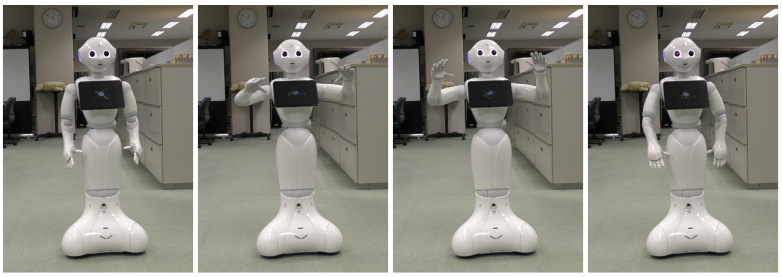
and the arms positioned along the body. Then, the right hand is gradually raised up while the left hand position remains unchanged as appeared in the initial position. This gesture is ended by putting down the right hand to the original pose. Overall, the sequence of frames on both the real and generated motion looks like a person is waving the right hand. However, the corresponding pair of poses on each individual frame is different. This result suggests that our G model does not simply memorize and reproduce the training data. Similarly, to synthesize with the annotation *"A person waves with the left hand"*, the real sample starts with putting the left hand in front of the chest, while the generated action begins with the initial position as in the previous example. In Fig. 6.30b, it can be clearly seen that the motion produced by our proposed network is similar to the training data. Both the real and generated action represent the movement of the left hand while the position of the right hand remains unchanged over time. For quantitative evaluation, again, we measured the average similarity between motions produced from proposed approaches and real ones using evaluation metrics as similar as applied in section 6.3. Table 6.12 presents the average similarity conducted in

Table 6.12: Average similarity between generated actions and ground truth actions: a comparison among 1 Channel [2], 3 Channel without Encoder/Decoder phase [2] (w/o E/D), and fully implemented model [3] (full model) approach.

	1 Channel	w/o E/D	full model
Average similarity	0.5931	0.6364	0.6603



(a)

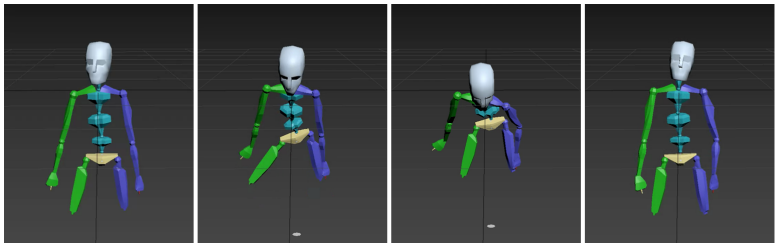


(b)

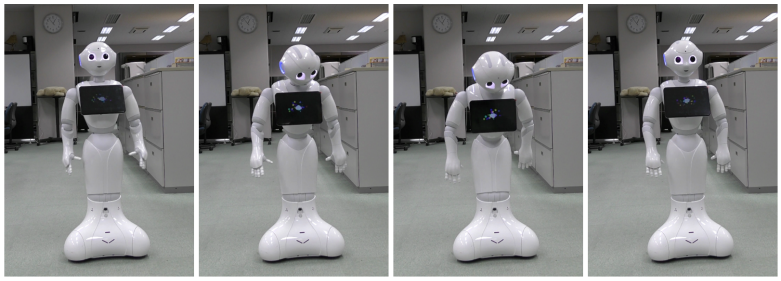
Figure 6.32: Fig. 6.32a shows the generated action by giving the input "someone over their is waving with their both two hands". Through the transformation model, the action is performed by the target robot as in Fig. 6.32b.

testing data. The experimental results, again, confirm that actions produced from the fully implemented model are more similar to the ground truth actions.

Fig. 6.32, and 6.33 show the actions synthesized with *"Someone over there is waving with their both two hands"* and *"They are taking a deep bow to show their respect"*. Noticed that those sentences are not available in the dataset. Instead, those speech texts were modified while keeping the original meaning of *"waving both hands"* and *"make a bow"*. The produced motion in Fig. 6.32a can be observed as someone wave with his or her two hands. Similarly, the result in Fig. 6.33a presents a sequence of frames as a person is collapsing their body downward while the arms are kept lower than the hip. Using the proposed transformation model illustrated in chapter 5, the generated human ac-



(a)



(b)

Figure 6.33: The generated action by feeding the input "they are taking a deep bow to show their respect".

tions were transformed into the Pepper robot's motion, representing by $\theta = \{\alpha_{Hip}, \beta_{Hip}, \beta_{Knee}, \alpha_{RightShoulder}, \beta_{RightShoulder}, \alpha_{RightElbow}, \gamma_{RightElbow}, \alpha_{LeftShoulder}, \beta_{LeftShoulder}, \alpha_{LeftElbow}, \gamma_{LeftElbow}, \beta_{Head}, \gamma_{Head}\}$. As displayed in Fig. 6.32b, the action performed by the Pepper robot preserves its original meaning as displayed on the human skeleton in Fig. 6.32a. In order to synchronize with the text description *"Someone over there is waving with their both two hands"*, from the initial pose, Pepper is gradually moving its two hands over the shoulder and then waving. Fig. 6.33b shows the generated robot's gesture from the proposed approach by filling the input *"They are taking a deep bow to show their respect"*. The result shows that the action looks like Pepper is collapsing its upper body while its two hands remained unchanged. It should be noted that when transforming human-like actions into Pepper, the robot's joint angles are checked with the joint boundary constraints before releasing. Thus, the generated motion displayed on the target robot in Fig. 6.33b shows that the Pepper could not bend their upper body as much as performed by the human skeleton in Fig. 6.33a. It can be observed in the robot's bodily expressions that at the same time with bending its hip, Pepper

also turns its head down as similar as displayed in the human action. Here, by additionally equipping the head's movement, the message "taking a deep bow" is more recognizable. Compared to the experiment conducted in section 6.3, in this experiment, robot actions have been defined by a higher number of DoF. This extension allows the target robot to execute generated actions in a more efficient way. As a result, sophisticated contexts of the robot's speech could be expressed in a transparent manner.

6.3.4 Summary

In this experiment, we demonstrated the validity of the framework for generating communicative gestures on public datasets. At the generation phase, the Generator receives a speech text represented by an embedding vector as an input. The generative network produces a co-speech gesture conveying the meaning of the input sentence. Taking into account the human behavioral studies mentioned in chapter 2, such gestures are known as *iconic* or *metaphoric* gestures. To verify this approach, a series of experiments was conducted on the generated co-speech actions. Firstly, it is indicated that the designed model could imitate human joint distribution from the training data, and generate human-like gestures supporting the context of input speech. An evaluation metric was established to quantitatively confirm the synthesis between input sentences and the corresponding generated actions. The comparative results with related works verified that our produced actions are more natural and similar to the real ones. Furthermore, by utilizing the transformation model illustrated in chapter 5, generated human-like actions were transformed into the target robot taking into account the robot's physical constraints, and associated with the robot's speech. The experimental results suggested that compared to the action produced by the robot's off-the-shelf module, the robot's gestures created by our approach are more appropriately fit the semantic contents of the robot's speech. Finally, it is confirmed that the generative framework does not merely memorize and reproduce training data. It is able to produce a variety of gestures expressing the same meaning of input sentences. This promising result would encourage robots to perform novel gestures over time to support a certain context of their speech during interactions.

Chapter 7

Conclusion

7.1 Dissertation Summary

Non-verbal behaviors have an indispensable role in human-human interaction. Being encoded by the influence of human social behaviors, the generation of non-verbal cues is a growing interest research topic in the social robotics domain. Although non-verbal modalities are powerful tools, allowing humans to interact in a facile and transparent manner. However, non-verbal behaviors are somewhat ambiguous and highly affected by individual personality, cultural background, and so on. Those factors influence the way how people express and interpret non-verbal behaviors. Not only in the context of human-human interaction, but that influence has also been observed in social human-robot interactions as discussed in chapter 2. Taking into account the aspect of behavior adaption when generating robots' behaviors, this dissertation suggests an alternative approach to create robots' social behaviors through imitating interacting partners. Specifically, our approach endows robots capable of learning from human behaviors, obtained through long-term interaction, in an unsupervised manner. This approach emphasizes the use of bodily expressions as a reliable channel for (1) conveying emotional states, and (2) supporting concrete and abstract contents of speech. Consequentially, the frameworks for generating emotional and communicative gestures were proposed in chapter 3 and 4, respectively. The two models produce gestures defined in human motion space, through the designed transformation model in chapter 5,

human-like gestures are transformed into the target robot, being robots' social gestures.

A series of experiments was conducted in chapter 6 to verify the effectiveness of the proposed frameworks. In the first experiment, the 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.

In the second experiment, the model of generating emotional gestures, 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 social behaviors are affected by many different factors such as individual personalities and cultural backgrounds.

In the third experiment, the model of generating communicative gestures was validated on public datasets. The experimental results indicated that this approach could imitate human joint distribution from the training data and generate neutral human-like gestures. We have established an evaluation metric to quantitatively verify generated gestures. Compared to related works, generated motions produced by our framework are more natural and similar to the real human actions. Indeed, by integrating with the transform model, gestured human-like gestures were transformed into the Pepper robot and associated with the robot's speech. The experimental results indicated that compared to the action produced by the robot's off-the-shelf module, gestures created by our approach are more appropriately fit the semantic contents of the robot's speech.

7.2 Contributions

In this dissertation, the proposed approach emphasizes the role of human behaviors towards generating robots' non-verbal behaviors through imitation learning. Generated robots' social gestures could be used in different contexts of interaction. This research's outcomes could positively contribute to the development of non-verbal cues for social robots. In particular, the ones without dedicated facial articulation such as NAO, Pepper, Romeo, and so on. The main contributions of this dissertation are:

The framework for generating emotional gestures: this model provides robots a capability of learning human affective behaviors in an unsupervised manner. The generated gestures are used to convey robots' emotional states. Since the output actions are human joint coordinates defined in Cartesian space. The other studies could inherit this framework to create emotional gestures for different humanoid robot platforms.

The framework for generating communicative gestures: this approach allows robots to learn relations between human gestures and speech. The output actions are used for supporting semantic contents of robots' speech. Similar to the above-mentioned model, with this framework, generated human-like actions are defined in Cartesian space. It is straightforward to utilize this approach for other studies in social robotics and other related domains.

The transformation model: this framework is employed to convert human actions to the Pepper robot's motion space, taking into account the robot's physical constraints. Without integration with the framework for generating emotional or communicative gestures mentioned above, this model can also be used as a standalone function. In this case, through a one-shot demonstration, this function allows non-robotics users to teach the Pepper robot new social skills supporting different scenarios of daily interaction.

Implementation: To demonstrate the proposed approach on the Pepper robots, several external modules were implemented and integrated with the robot's built-in modules towards strengthening the robot's functionalities. Those extensions enable the Pepper robot a capability of entering into different interaction scenarios for collecting users' behaviors in an efficient manner. Indeed, the data of human

affective behaviors collected from the robot's point of view in our experiments could be used for other studies in the field of gesture recognition or generation.

7.3 Future Research Directions

Despite the given contributions, the work presented in this dissertation is just a small step towards enhancing the quality of social human-robot interaction through re-configurations of robots' nonverbal cues. Several interesting directions should be explored in the future to strengthen and extend the current work, some of them are the followings:

The proposed transformation model addresses the inverse kinematics problem based on geometric algebra. However, the exponential growth of social robots leads to the need of considering a more abstract approach, minimizing the workload for analytical modeling of the transformation process. Indeed, social robots are not necessarily designed in humanoid forms. It is interesting to explore how human social behaviors could be transformed into such robots while ensuring the human perception of robots' social cues. It is suggested that CycleGAN [134], and other similar techniques should be investigated for this transformation process.

In the designed model for generating emotional gestures, emotional bodily expressions are represented by discrete categories (*happiness*, *sadness*, *anger*, and so on). However, emotions can also be represented in affect space (e.g. Circumplex model of affect [124], PAD emotion model [64]). For instance, an emotion can be represented on a two-dimensional surface, including Arousal (ranging from deactivation to activation) and Valence (displeasure to pleasure). Using this approach, generated emotional gestures can be defined by continuous input values rather than discrete ones.

Concerning the framework for generating communicative gestures, our current work focuses on capturing relations between human gestures and the semantic contents of human speech. Generated gestures are employed to support concrete or abstract meanings of communicators' speech. Taking into account theories of human behaviors [7], those gestures are known as *iconic* and *metaphoric* gestures. However, the last two types of gestures known as *deictic*, and *beat* have not been investigated yet. It is important to remark that *deictic* gestures (e.g. pointing

gestures) are heavily connected to environmental information surrounding communicators. On the other hand, beat gestures (e.g. rhythm movements of hands) are correlated to communicators' speech prosody (audio features), rather than semantic contents. As the result, in order to enhance the gesture diversity towards supporting various contexts of communication, further signals should be equipped as the inputs to the current framework

Finally, in this study, emotional gestures are applied to express robots' emotional states, while communicative gestures are employed to support robots' speech. It is interesting to explore the combination of emotional and communicative gestures. In other words, robots' non-verbal behaviors should be able to support the contents of their speech, at the same time, signal their emotional states to the interacting partners. A possible approach to address this goal could be to manipulate several features (speed, amplitude, emotion, and so on) of generated communicative gestures while keeping action identity remained unchanged.

Appendix: Implementation of Observation Modules on the Pepper robot for Collecting Interacting Partners' Data

The experiments illustrated in chapter 6 was conducted with the Softbank humanoid Pepper robot. During experiments, most of the robot's off-the-shelf modules[1] have been utilized for demonstrating the proposed approach. Additionally, several external modules have been implemented and integrated with the robot's built-in modules. The main reason is that either because the function was not available on the robot, or the efficiency of that function was not meet our requirements for experimental setups. In this Appendix, we explain our practical implementations of the two extension modules: human pose estimation and human facial expression estimation.

Human Pose Estimation

Overview

In human action recognition and generation domains, it is common to present human actions as sequences of skeleton frames (known as motion data). Instead of presenting actions as raw input images, motion data requires less amount of mem-

[1] `http://doc.aldebaran.com/2-5/index_dev_guide.html`

ory while features of action are better presented in the form of skeleton frames. Here, a human skeleton frame is a schematic model of the locations of the torso, head, and limb of a human body. Overall, there are two main approaches for estimating a sequence of skeleton frames from a human action: marker sensors and markerless sensors. For collecting interacting partners' information through scenarios of social human-robot interaction, rather than equipping the users with a set of marker sensors, we have decided to use the markerless sensors based approach. The main reason is that the marker sensors based approach requires complicated setups before conducting an interaction section. Indeed, the interacting partners may feel uncomfortable when equipping external sensors on their bodies during social interactions. Those reasons suggest that the use of marker sensors based approach may reduce the quality of interaction. Vice versa, with the markerless sensors based approach, external cameras are utilized for capturing the interacting partners' actions and presenting that motion as sequences of skeleton frames. Thus, this approach would not influence scenarios of interaction, especially, when such cameras are embedded in the target robot. The following parts will explain our implementation in more detail.

Implementation on the Target Robot

For collecting users' motion data using the markerless sensors based approach, the Microsoft Kinect sensor has been used in our preliminary works. The accuracy of human pose estimation from this sensor was acceptable for our requirements of experimental setups. However, taking into account the scenario of day-to-day human-robot interaction, rather than using external devices that cause the data-privacy issue, it is recommended that robots should be able to perceive environmental stimuli by their on-board sensors. To tackle this problem, different solutions have been tested and evaluated by taking into account the available sensors[2] embedded in the Pepper robot. First of all, as an unofficial function provided for the robot[3], this module collects images captured from the robot's Asus Xtion camera[4]. The experiments noticed that due to limitations of the robot's com-

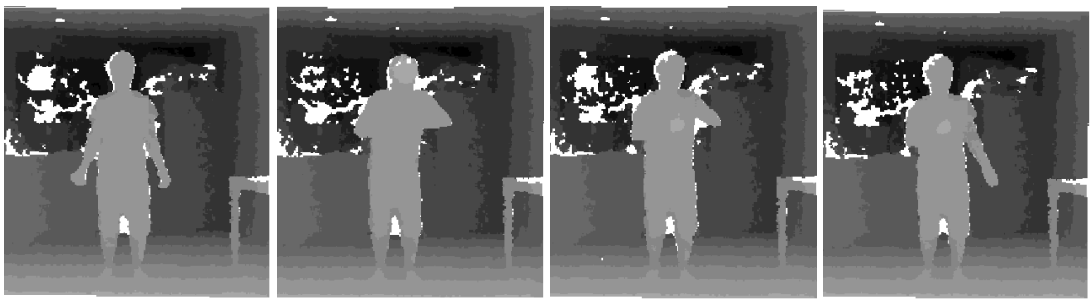
[2] http://doc.aldebaran.com/2-5/family/pepper_technical/video_overview.html
[3] http://protolab.aldebaran.com:9000/protolab/SkeletonDetector
[4] http://doc.aldebaran.com/2-5/family/pepper_technical/video_3D_pep.html

putation resources, several functionalities of the Asus Xtion sensor had been cut down or modified when embedding into the target robot. Such modifications significantly reduced the accuracy of the pose estimation module. Another approach is that utilizing the images captured from the robot's 2D camera[5] for estimating human pose [121]. It is reported that the performance of skeleton estimation is qualitatively comparable with other monocular RGB-D sensors based approach such as Kinect or Asus Xtion. Thirdly, both RGB and depth images obtained from the robot's on-board module were fed to pose estimation mechanism [135]. It is started by extracting a 2D skeleton frame from an RGB image. The 2D pose is then combined with the associated depth image for estimating the 3D human pose. Except for the first approach which shows the low performance as discussed above, the last two approaches were implemented in the target robot as the human pose estimation modules. Figs. 7.17.27.3 show an example case of human RGB images, the corresponding depth images and the ground truth skeleton, respectively. Fig. 7.4 presents the estimated pose by receiving RGB images as the inputs while Fig. 7.5 displays the results of the pose estimation mechanism by feeding both RGB and depth images. Finally, the differences between the ground truth pose and the estimated poses are qualitatively illustrated in Fig. 7.6. Fig. 7.7 demonstrates the operation of human pose estimation module in the scenario of human-robot interaction conducted in Experiment 6.1.2 of chapter 6.

[5] `http://doc.aldebaran.com/2-5/family/pepper_technical/video_2D_pep.html`

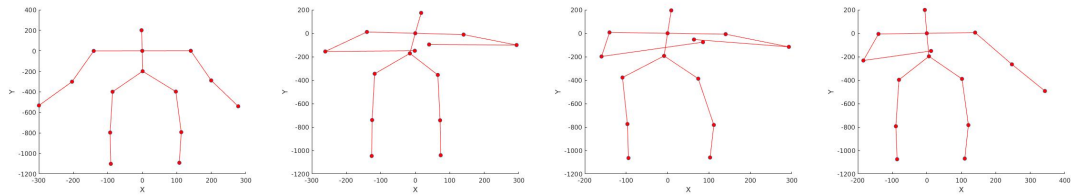
(a) (b) (c) (d)

Figure 7.1: The RGB images of the demonstrator.



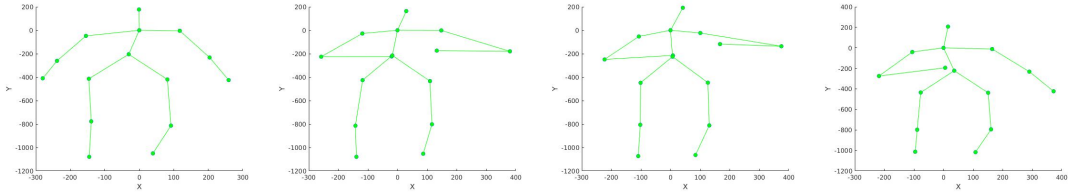
(a) (b) (c) (d)

Figure 7.2: The depth images of the demonstrator.



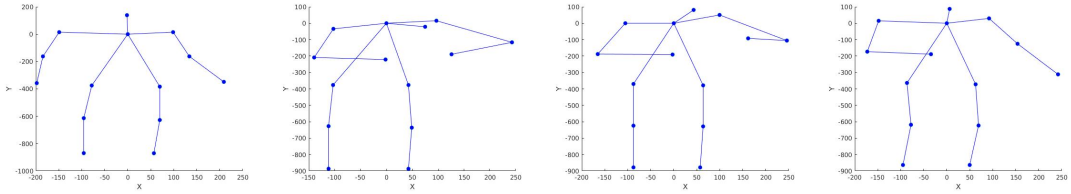
(a) (b) (c) (d)

Figure 7.3: The ground truth skeleton of the demonstrator.



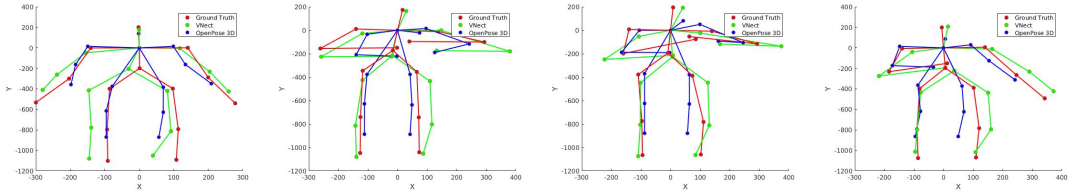
(a) (b) (c) (d)

Figure 7.4: The estimated skeleton from the RGB images.



(a) (b) (c) (d)

Figure 7.5: The estimated skeleton from the combination between depth and RGB images.

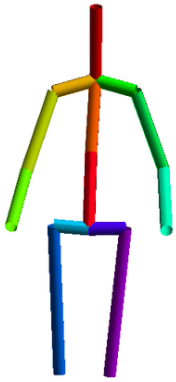


(a) (b) (c) (d)

Figure 7.6: The differences between estimated the skeleton frames and the ground truth ones.

(a) The user stands in front Pepper robot for social interaction

(b) The user image captured from the robot's filed of view



(c) The user image captured from the robot's filed of view

Figure 7.7: Operation of the pose estimation module in a scenario of interaction.

Human Facial Expression Estimation

Overview

By estimating the users' emotions, robots can assess the effects of their behaviors and modify them adapting to the users' mental states. Emotions can be recognized through a variety of means such as voice intonations, body language, facial movements, or electroencephalograpy (EEG). Among these modalities, the face channel is one of the primary means for conveying human emotions, leading thereby to a practical approach to recognizing the user's emotion. Taking into theories of human emotion [136], emotions could be categorized into several basic groups, those are recognized across a wide range of cultures such as anger, disgust, fear, happiness, sadness, surprise, and contempt. Several scenarios of social interactions discussed in chapter 6 requires robots capable of estimating user emotions through facial expressions. In addition to the robot's off-the-shelf module, different approaches using publicly available APIs have been implemented, allowing the Pepper robot to estimate the interacting partners' emotions from their facial expressions in an efficient manner.

Implementation on the Target Robot

It is noticed that interacting partners' emotions could be estimated by using the robot off-the-shelf module ALFaceCharacteristic[6]. This API releases an array of the confidence value of 5 expressions: *neutral, happy, surprised, angry*, and *sad*. Additionally, the Microsoft Azure Emotion API[7] and the Kairos Emotion Analysis API[8] were implemented on the target robot. The Microsoft Azure API takes a human facial expression as an image input and returns the confident value of *anger, contempt, disgust, fear, happiness, neutral, sadness*, and *surprise* as an output array. In practice, the NAOqi ALPeoplePerception API[9] is firstly subscribed. Once an interacting partner is detected, the user's facial expression is sent to one of the available APIs as an emotion estimation request and receives emotional values as a response.

In order to quantitatively evaluate performances of APIs (Kairos Emotion Analysis API, Microsoft Azure Emotion API, and the NAOqi ALFaceCharacteristic API) for the user's facial expressions, an experiment was conducted on the Karolinska Directed Emotional Faces (KDEF) public dataset [137]. KDEF contains 7 different emotions: afraid, angry, disgust, happy, neutral, sad, and surprise of 140 subjects. Firstly, the experiment was carried out with the Kairos API whose performance was described in Table 7.1. Next, the Microsoft Azure API was tested on the same dataset. This API classifies human facial expressions into 8 different labels: *anger, contempt, disgust, fear, happiness, neutral, sadness*, and *surprise*. It was noticed that the emotional label "contempt" was not available in the KDEF dataset. Therefore, any images estimated as "contempt" were ignored. The performance is summarized in Table 7.2. Finally, the NAOqi *ALFaceCharacteristic* API was tested on the KDEF dataset. The facial images from the KDEF dataset were sequentially presented to Pepper. Specifically, the facial images detected by the *ALPeoplePerception* API were analyzed by the *ALFaceCharacteristic* API. Finally, the corresponding emotion was received from the *ALFaceCharacteristic* API that

[6] `http://doc.aldebaran.com/2-5/naoqi/peopleperception/alfacecharacteristics.html#alfacecharacteristics`

[7] `https://azure.microsoft.com/en-us/services/cognitive-services/emotion/`

[8] `https://www.kairos.com/emotion-analysis-api`

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

returns an array of the detection score of five expressions: *neutral, happy, surprised, angry*, and *sad*. In other words, this test was conducted on the KDEF images with 5 such emotion labels. The performance of *ALFaceCharacteristic* was summarized as the confusion matrix illustrated in Table 7.3.

Table 7.1: Precision, Recall and F1-score of Kairos API with KDEF dataset

	Precision	**Recall**	**F1-score**
joy	0.70	0.62	0.66
surprise	0.78	0.53	0.63
disgust	0.72	0.69	0.70
sadness	0.71	0.73	0.72
anger	0.69	0.79	0.74
fear	0.82	0.91	0.86
Average	0.74	0.74	0.73

Table 7.2: Precision, Recall and F1-score of Microsoft Azure API with KDEF dataset

	Precision	**Recall**	**F1-score**
happiness	0.86	0.57	0.68
surprise	0.96	0.71	0.82
fear	0.96	0.18	0.30
neutral	0.93	1.00	0.96
disgust	0.60	1.00	0.75
anger	0.70	0.86	0.77
sadness	0.68	0.96	0.80
Average	0.81	0.76	0.73

Table 7.3: Precision, Recall and F1-score of NAOqi ALPeoplePerception with KDEF dataset

	Precision	**Recall**	**F1-score**
angry	0.48	0.74	0.58
happy	0.55	0.53	0.54
neutral	0.53	0.53	0.53
sad	0.48	0.36	0.41
surprised	0.68	0.51	0.58
Average	0.54	0.53	0.53

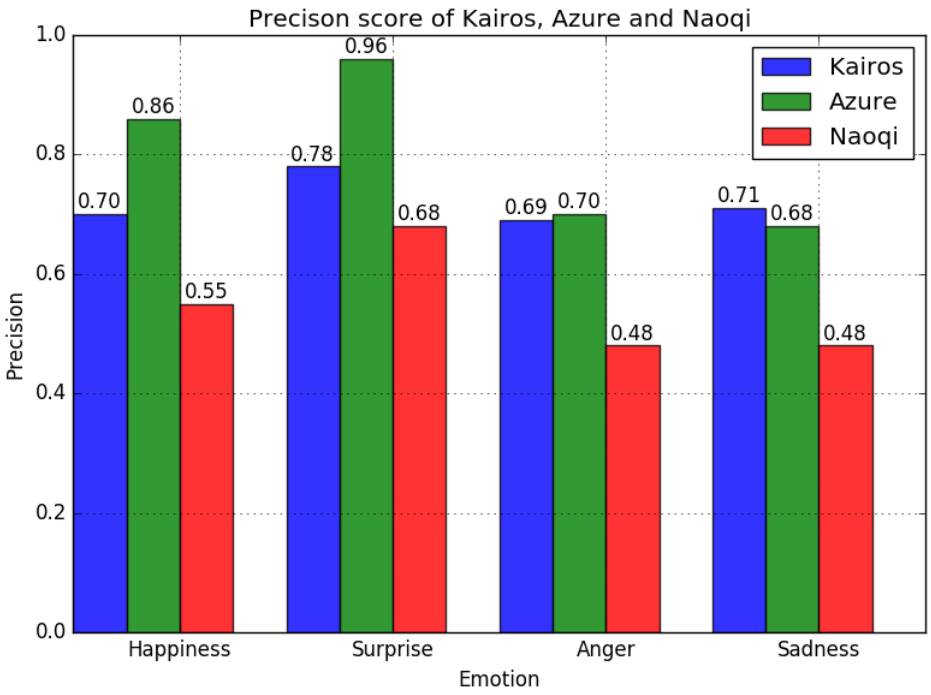


Figure 7.8: Performance of Kairos, Microsoft Auzre, and NAOqi for emotion estimation task.

It is difficult to make a fair comparison of Kairos, Microsoft Azure, and NAOqi *ALFaceCharacteristic* with the KDEF dataset due to the differences in labeling estimated emotions among those approaches, as addressed before. Therefore, the score of precision of 4 emotion labels *Happiness, Surprise, Anger, and Sadness* that commonly available among the 3 APIs was considered. It is clear from Fig. 7.8 that the Microsoft Azure API shows the best performance in most cases. It is also noted that the Naoqi *ALFaceCharacteristic* shows the best performance in terms of processing time since it does not require any communication overhead. Finally, this comparison can only be considered as a reference due to the difference in emotion labels across different APIs and a different experimental setup for the NAOqi API as described above which also affects the performance of *ALFaceCharacteristic*.

Fig. 7.9 demonstrates an example case of estimating users' emotions through facial expression. Noticed that, this module can be used as a stand-alone function. In that case, the estimated emotions are displayed on the robot table as emojis.

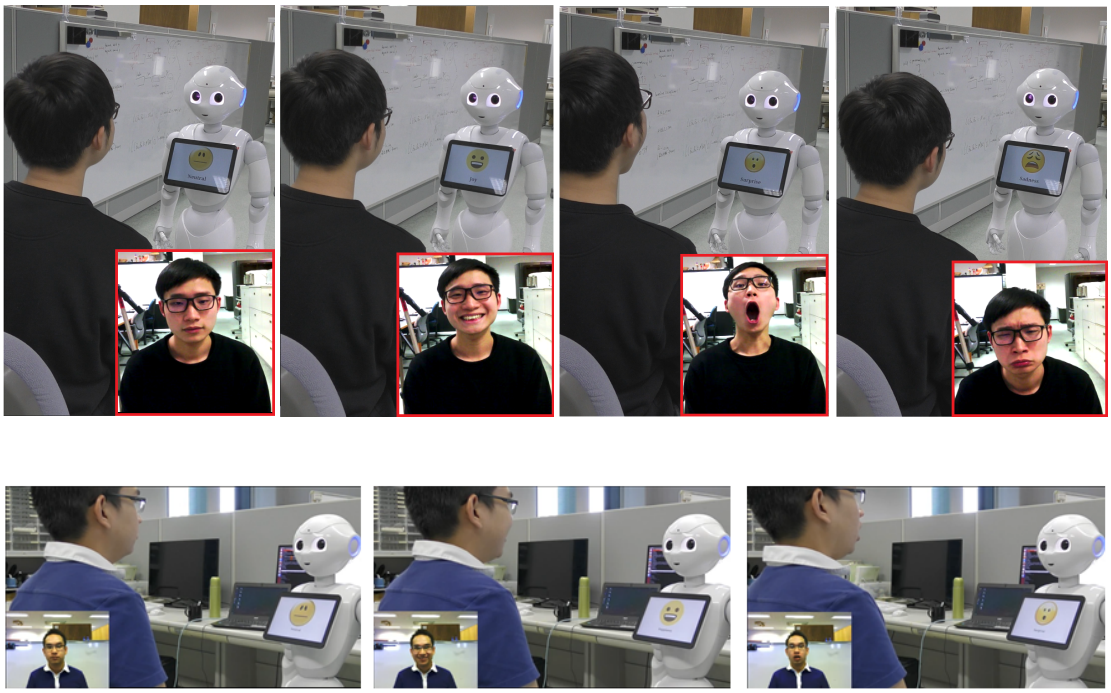


Figure 7.9: The robot tracks human facial expression. The estimated emotion is displayed on Pepper's tablet. The small boxes indicate images from the robot's field of view.

On the other hand, by integrating this module with the human pose estimation mentioned in section 7.3, the Pepper robot could collect the user's social behaviors through a scenario of interaction as discussed in section 6.2 of chapter 6.

Bibliography

[1] H. Ahn, T. Ha, Y. Choi, H. Yoo, and S. Oh, "Text2action: Generative adversarial synthesis from language to action," in *IEEE Int'l Conf. on Robotics and Automation*, pp. 5915–5920, 2018.

[2] N. T. V. Tuyen, A. Elibol, and N. Y. Chong, "Learning from humans to generate communicative gestures for social robots," in *2020 17th International Conference on Ubiquitous Robots (UR)*, pp. 284–289, IEEE, 2020.

[3] N. T. V. Tuyen, A. Elibol, and N. Y. Chong, "Conditional generative adversarial network for generating communicative robot gestures," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 201–207, IEEE, 2020.

[4] P. Ekman and W. Friesen, "A technique for the measurement of facial movement," *Facial Action Coding System*, 1978.

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

[6] H. G. Wallbott, "Bodily expression of emotion," *European journal of social psychology*, vol. 28, no. 6, pp. 879–896, 1998.

[7] D. McNeill, *Hand and mind: What gestures reveal about thought*. University of Chicago press, 1992.

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

[9] 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, pp. 204–209, IEEE, 2011.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] R. W. Picard and W. Rosalind, "Toward agents that recognize emotion," *VIVEK-BOMBAY-*, vol. 13, no. 1, pp. 3–13, 2000.

[16] 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 *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pp. 1–8, IEEE, 2019.

[17] 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.

[18] A. Aly and A. Tapus, "Towards an intelligent system for generating an adapted verbal and nonverbal combined behavior in human–robot interaction," *Autonomous Robots*, vol. 40, no. 2, pp. 193–209, 2016.

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

[20] S. Robotics, "Pepper robot." `http://doc.aldebaran.com/2-5/naoqi/motion/alanimationplayer-advanced.html#animationplayer-list-behaviors-pepper`. Accessed: 2020-11-17.

[21] S. Robotics, "Nao robot." `http://doc.aldebaran.com/2-5/naoqi/motion/alanimationplayer-advanced.html#animationplayer-list-behaviors-nao`. Accessed: 2020-11-17.

[22] J. Cassell, H. H. Vilhjálmsson, and T. Bickmore, "Beat: the behavior expression animation toolkit," in *Life-Like Characters*, pp. 163–185, Springer, 2004.

[23] V. Ng-Thow-Hing, P. Luo, and S. Okita, "Synchronized gesture and speech production for humanoid robots," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 4617–4624, IEEE, 2010.

[24] C.-M. Huang and B. Mutlu, "Robot behavior toolkit: generating effective social behaviors for robots," in *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 25–32, IEEE, 2012.

[25] B. Ravenet, C. Pelachaud, C. Clavel, and S. Marsella, "Automating the production of communicative gestures in embodied characters," *Frontiers in psychology*, vol. 9, p. 1144, 2018.

[26] J. Xu, J. Broekens, K. Hindriks, and M. A. Neerincx, "Effects of a robotic storyteller's moody gestures on storytelling perception," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 449–455, IEEE, 2015.

[27] I. Ranatunga, M. Beltran, N. A. Torres, N. Bugnariu, R. M. Patterson, C. Garver, and D. O. Popa, "Human-robot upper body gesture imitation analysis for autism spectrum disorders," in *International Conference on Social Robotics*, pp. 218–228, Springer, 2013.

[28] A. Adams and P. Robinson, "An android head for social-emotional intervention for children with autism spectrum conditions," in *International Conference on Affective Computing and Intelligent Interaction*, pp. 183–190, Springer, 2011.

[29] Y. Kondo and Y. Takahashi, "Real-time whole body imitation by humanoic robot based on particle filter and dimension reduction by autoencoder," in *2017 Joint 17th World Congress of International Fuzzy Systems Association and 9th International Conference on Soft Computing and Intelligent Systems (IFSA-SCIS)*, pp. 1–6, IEEE, 2017.

[30] H. Guedjou, S. Boucenna, J. Xavier, D. Cohen, and M. Chetouani, "The influence of individual social traits on robot learning in a human-robot interaction," in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 256–262, IEEE, 2017.

[31] P. Bremner and U. Leonards, "Iconic gestures for robot avatars, recognition and integration with speech," *Frontiers in psychology*, vol. 7, p. 183, 2016.

[32] C. Frith, "Role of facial expressions in social interactions," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1535, pp. 3453–3458, 2009.

[33] F. Noroozi, D. Kaminska, C. Corneanu, T. Sapinski, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *IEEE transactions on affective computing*, 2018.

[34] G.-B. Duchenne and G.-B. D. de Boulogne, *The mechanism of human facial expression*. Cambridge university press, 1990.

[35] H. Ruthrof, *The body in language*. Bloomsbury Publishing, 2015.

[36] J. Van den Stock, R. Righart, and B. De Gelder, "Body expressions influence recognition of emotions in the face and voice.," *Emotion*, vol. 7, no. 3, p. 487, 2007.

[37] B. De Gelder, "Why bodies? twelve reasons for including bodily expressions in affective neuroscience," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 364, no. 1535, pp. 3475–3484, 2009.

[38] J. Montepare, E. Koff, D. Zaitchik, and M. Albert, "The use of body movements and gestures as cues to emotions in younger and older adults," *Journal of Nonverbal Behavior*, vol. 23, no. 2, pp. 133–152, 1999.

[39] H. K. Meeren, C. C. van Heijnsbergen, and B. de Gelder, "Rapid perceptual integration of facial expression and emotional body language," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 45, pp. 16518–16523, 2005.

[40] A. Mehrabian and J. T. Friar, "Encoding of attitude by a seated communicator via posture and position cues.," *Journal of Consulting and Clinical Psychology*, vol. 33, no. 3, p. 330, 1969.

[41] A. B. Hostetter, "When do gestures communicate? a meta-analysis.," *Psychological bulletin*, vol. 137, no. 2, p. 297, 2011.

[42] S. Goldin-Meadow, "The role of gesture in communication and thinking," *Trends in cognitive sciences*, vol. 3, no. 11, pp. 419–429, 1999.

[43] S. Goldin-Meadow, *Hearing gesture: How our hands help us think*. Harvard University Press, 2005.

[44] 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.

[45] C. Breazeal, "Emotion and sociable humanoid robots," *International journal of human-computer studies*, vol. 59, no. 1-2, pp. 119–155, 2003.

[46] D. Cameron, A. Millings, S. Fernando, E. C. Collins, R. Moore, A. Sharkey, V. Evers, and T. Prescott, "The effects of robot facial emotional expressions and gender on child–robot interaction in a field study," *Connection science*, vol. 30, no. 4, pp. 343–361, 2018.

[47] T. L. Q. Dang, N. T. V. Tuyen, S. Jeong, and N. Y. Chong, "Encoding cultures in robot emotion representation," in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 547–552, IEEE, 2017.

[48] C.-M. Huang and B. Mutlu, "Modeling and evaluating narrative gestures for humanlike robots.," in *Robotics: Science and Systems*, pp. 57–64, 2013.

[49] C. Stanton and C. J. Stevens, "Robot pressure: the impact of robot eye gaze and lifelike bodily movements upon decision-making and trust," in *International Conference on Social Robotics*, pp. 330–339, Springer, 2014.

[50] P. Patompak, S. Jeong, I. Nilkhamhang, and N. Y. Chong, "Learning proxemics for personalized human–robot social interaction," *International Journal of Social Robotics*, pp. 1–14, 2019.

[51] 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.

[52] K. Bergmann and M. Macedonia, "A virtual agent as vocabulary trainer: iconic gestures help to improve learners' memory performance," in *International workshop on intelligent virtual agents*, pp. 139–148, Springer, 2013.

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

[54] 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.

[55] 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 *2017 26th IEEE international symposium on robot and human interactive communication (RO-MAN)*, pp. 553–560, IEEE, 2017.

[56] 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*, pp. 15–54, Springer, 1992.

[57] S. Feinman, "Social referencing in infancy," *Merrill-Palmer Quarterly (1982-)*, pp. 445–470, 1982.

[58] A. N. Meltzoff, "The role of imitation in understanding persons and developing a theory of mind," *Understanding other minds: Perspectives from autism*, 1993.

[59] R. Hortensius, F. Hekele, and E. S. Crossy, "The perception of emotion in artificial agents," *IEEE Transactions on Cognitive and Developmental Systems*, 2018.

[60] 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.

[61] 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.

[62] R. T. Boone and J. G. Cunningham, "Children's expression of emotional meaning in music through expressive body movement," *Journal of nonverbal behavior*, vol. 25, no. 1, pp. 21–41, 2001.

[63] M. Coulson, "Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence," *Journal of nonverbal behavior*, vol. 28, no. 2, pp. 117–139, 2004.

[64] A. Mehrabian and J. A. Russell, *An approach to environmental psychology.* the MIT Press, 1974.

[65] A. Beck, *Perception of emotional body language displayed by animated characters.* PhD thesis, University of Portsmouth, 2011.

[66] 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*, pp. 50–59, Springer, 2005.

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

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

[69] 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.

[70] 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.

[71] S. Valipour, C. Perez, and M. Jagersand, "Incremental learning for robot perception through hri," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2772–2777, IEEE, 2017.

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

[73] 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 *Int'l Joint Conf. on Artificial Intelligence*, vol. 13, pp. 2466–2472, 2013.

[74] 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.

[75] N. T. V. Tuyen, S. Jeong, and N. Y. Chong, "Incremental learning of human emotional behavior for social robot emotional body expression," in *2018 15th International Conference on Ubiquitous Robots (UR)*, pp. 377–382, IEEE, 2018.

[76] N. T. V. Tuyen, S. Jeong, and N. Y. Chong, "Learning human behavior for emotional body expression in socially assistive robotics," in *Ubiquitous Robots and Ambient Intelligence (URAI), 2017 14th International Conference on*, pp. 45–50, IEEE, 2017.

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

[78] H. Fang, Y. Du, L. Xia, J. Li, J. Zhang, and K. Wang, "A topology-preserving selection and clustering approach to multidimensional biological data," *Omics: a journal of integrative biology*, vol. 15, no. 7-8, pp. 483–494, 2011.

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

[80] I. Ahrns, J. Bruske, and G. Sommer, "On-line learning with dynamic cell structures," in *Proceedings of the International Conference on Artificial Neural Networks*, vol. 2, pp. 141–146, 1995.

[81] T. Martinetz, "Competitive hebbian learning rule forms perfectly topology preserving maps," in *International conference on artificial neural networks*, pp. 427–434, Springer, 1993.

[82] B. Fritzke, "Growing cell structures—a self-organizing network for unsupervised and supervised learning," *Neural networks*, vol. 7, no. 9, pp. 1441–1460, 1994.

[83] B. Fritzke, "A growing neural gas network learns topologies," in *Advances in neural information processing systems*, pp. 625–632, 1995.

[84] 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.

[85] N. Elfaramawy, P. Barros, G. I. Parisi, and S. Wermter, "Emotion recognition from body expressions with a neural network architecture," in *Proceedings of the 5th International Conference on Human Agent Interaction*, pp. 143–149, 2017.

[86] M. G. Perhinschi, G. Campa, M. R. Napolitano, M. Lando, L. Massotti, and M. L. Fravolini, "A simulation tool for on-line real time parameter identification," in *Proceedings of the 2002 AIAA modeling and simulation conference*, 2002.

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

[88] J. Lampinen and E. Oja, "Clustering properties of hierarchical self-organizing maps," in *Mathematical Nonlinear Image Processing*, pp. 165–176, Springer, 1993.

[89] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 224–227, 1979.

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

[91] T. Kucherenko, "Data driven non-verbal behavior generation for humanoid robots," in *International Conference on Multimodal Interaction*, pp. 520–523, ACM, 2018.

[92] S. Marsella, Y. Xu, M. Lhommet, A. Feng, S. Scherer, and A. Shapiro, "Virtual character performance from speech," in *ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pp. 25–35, 2013.

[93] C. L. Breazeal, *Designing sociable robots*. MIT press, 2002.

[94] P. Ekman, "Are there basic emotions?," *Psychological Review*, 1992.

[95] A. Shimazu, C. Hieida, T. Nagai, T. Nakamura, Y. Takeda, T. Hara, O. Nakagawa, and T. Maeda, "Generation of gestures during presentation for humanoid robots," in *IEEE Int'l Symposium on Robot and Human Interactive Communication*, pp. 961–968, 2018.

[96] T. Kucherenko, D. Hasegawa, G. E. Henter, N. Kaneko, and H. Kjellström, "Analyzing input and output representations for speech-driven gesture generation," in *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pp. 97–104, 2019.

[97] N. Sadoughi and C. Busso, "Joint learning of speech-driven facial motion with bidirectional long-short term memory," in *International Conference on Intelligent Virtual Agents*, pp. 389–402, Springer, 2017.

[98] Y. Yoon, W.-R. Ko, M. Jang, J. Lee, J. Kim, and G. Lee, "Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots," *arXiv preprint arXiv:1810.12541*, 2018.

[99] P. Wolfert, T. Kucherenko, H. Kjelström, and T. Belpaeme, "Should beat gestures be learned or designed? a benchmarking user study," in *ICDL-EPIROB 2019 Workshop on Naturalistic Non-Verbal and Affective Human-Robot Interactions*, 2019.

[100] M. Plappert, C. Mandery, and T. Asfour, "Learning a bidirectional mapping between human whole-body motion and natural language using deep

recurrent neural networks," *Robotics and Autonomous Systems*, vol. 109, pp. 13–26, 2018.

[101] K. Wang, C. Gou, Y. Duan, Y. Lin, X. Zheng, and F.-Y. Wang, "Generative adversarial networks: Introduction and outlook," *IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 4, pp. 588–598, 2017.

[102] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[103] J. Gauthier, "Conditional generative adversarial nets for convolutional face generation," *Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester*, vol. 2014, no. 5, p. 2, 2014.

[104] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," in *Advances in Neural Information Processing Systems*, pp. 613–621, 2016.

[105] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," *arXiv preprint arXiv:1802.04208*, 2018.

[106] Y. Yan, J. Xu, B. Ni, W. Zhang, and X. Yang, "Skeleton-aided articulated motion generation," in *ACM International Conference on Multimedia*, pp. 199–207, ACM, 2017.

[107] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[108] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler, "Skip-thought vectors," in *Advances in Neural Information Processing Systems*, pp. 3294–3302, 2015.

[109] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[110] Z. Yang, Y. Li, J. Yang, and J. Luo, "Action recognition with spatio–temporal visual attention on skeleton image sequences," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 8, pp. 2405–2415, 2018.

[111] D. Holden, J. Saito, and T. Komura, "A deep learning framework for character motion synthesis and editing," *ACM Transactions on Graphics*, vol. 35, no. 4, pp. 1–11, 2016.

[112] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 3288–3297, 2017.

[113] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Int'l Conf. on Machine Learning*, pp. 807–814, 2010.

[114] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robotics and autonomous systems*, vol. 57, no. 5, pp. 469–483, 2009.

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

[116] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng, "Ros: an open-source robot operating system," in *ICRA workshop on open source software*, vol. 3, p. 5, Kobe, Japan, 2009.

[117] S. Cotton, M. Vanoncini, P. Fraisse, N. Ramdani, E. Demircan, A. P. Murray, and T. Keller, "Estimation of the centre of mass from motion capture and force plate recordings: a study on the elderly," *Applied Bionics and Biomechanics*, vol. 8, no. 1, pp. 67–84, 2011.

[118] J. Koenemann, F. Burget, and M. Bennewitz, "Real-time imitation of human whole-body motions by humanoids," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2806–2812, IEEE, 2014.

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

[120] C. Stanton, A. Bogdanovych, and E. Ratanasena, "Teleoperation of a humanoid robot using full-body motion capture, example movements, and machine learning," in *Proc. Australasian Conference on Robotics and Automation*, vol. 8, p. 51, 2012.

[121] 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.

[122] 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.

[123] 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.

[124] 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.

[125] 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.

[126] 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*, pp. 213–222, 2017.

[127] 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.

[128] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5288–5296, 2016.

[129] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *IEEE Int'l Conf. on Computer Vision*, pp. 19–27, 2015.

[130] N. T. V. Tuyen, A. Elibol, and N. Y. Chong, "Learning bodily expression of emotion for social robots through human interaction," *IEEE Transactions on Cognitive and Developmental Systems*, 2020.

[131] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[132] C. Mandery, Ö. Terlemez, M. Do, N. Vahrenkamp, and T. Asfour, "The kit whole-body human motion database," in *International Conference on Advanced Robotics*, pp. 329–336, IEEE, 2015.

[133] M. Plappert, C. Mandery, and T. Asfour, "The kit motion-language dataset," *Big Data*, vol. 4, no. 4, pp. 236–252, 2016.

[134] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

[135] C. Zimmermann, T. Welschehold, C. Dornhege, W. Burgard, and T. Brox, "3d human pose estimation in rgbd images for robotic task learning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1986–1992, IEEE, 2018.

[136] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion.," *Journal of personality and social psychology*, vol. 17, no. 2, p. 124, 1971.

[137] D. Lundqvist, A. Flykt, and A. Öhman, "The karolinska directed emotional faces (kdef)," *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet*, vol. 91, no. 630, pp. 2–2, 1998.

Publications

Journal

[1] N.T.V.Tuyen, A.Elibol, and N.Y.Chong. "Learning Bodily Expression of Emotion for Social Robots through Human Interaction," in *IEEE Transactions on Cognitive and Developmental Systems*, vol.13, no.1, pp.16-30, March 2021.

International Conference

[1] N.T.V.Tuyen, A.Elibol, and N.Y.Chong. "Conditional Generative Adversarial Network for Generating Communicative Robot Gesture," in *2020 IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*, pp.201-207, IEEE, 2020.

[2] N.T.V.Tuyen, A.Elibol, and N.Y.Chong. "Learning from Humans to Generate Communicative Gestures for Social Robots," in *2020 International Conference on Ubiquitous Robots (UR)*, pp.284-289, IEEE, 2020.

[3] N.T.V.Tuyen, S.Jeong, and N.Y.Chong. "Emotional Bodily Expressions for Culturally Competent Robots through Long Term Human-robot Interaction," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp.2008-2013, IEEE, 2018.

[4] N.T.V.Tuyen, S.Jeong, and N.Y.Chong. "Incremental Learning of Human Emotional Behavior for Social Robot Emotional Body Expression," in *2018*

International Conference on Ubiquitous Robots (UR), pp.377-382, IEEE, 2018.

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

[6] TLQ Dang, N.T.V.Tuyen, S.Jeong, and N.Y.Chong. "Encoding cultures in robot emotion representation," in *2017 IEEE International Symposium on Robot and Human Interactive Communication (ROMAN)*, pp.547-552, IEEE, 2017.

International Workshop

[1] N.T.V.Tuyen, A.Elibol, and N.Y.Chong. "Generating Social Behaviors through Imitation Learning for Socially Assistive Robots," in *2019 JAIST Smart Information, Smart Knowledge, Smart Material Workshop (SISKSM)*, 2019.