"Can You Guess My Moves?" Playing Charades with a Humanoid Robot Employing Mutual Learning with Emotional Intelligence

Baijun Xie
bdxie@gwu.edu
Department of Biomedical Engineering
School of Engineering and Applied Science
George Washington University
Washington, D.C., USA

Chung Hyuk Park
chpark@gwu.edu
Department of Biomedical Engineering
School of Engineering and Applied Science
George Washington University
Washington, D.C., USA

ABSTRACT

Social play is essential in human interactions, increasing social bonding, mitigating stress, and relieving anxiety. With advancements in robotics, social robots can employ this role to assist in human-robot interaction scenarios for clinical and healthcare purposes. However, robotic intelligence still needs further development to match the wide spectrum of social behaviors and contexts in human interactions. In this paper, we present our robotic intelligence framework with a mutual learning paradigm in which we apply deep learning based on emotion recognition and behavior perception, through which the robot learns human movements and contexts through the interactive game of charades. Furthermore, we designed a gesture-based social game to provide a more empathetic and engaging social robot for the user. We also created a custom behavior database containing contextual behaviors for the proposed social games. A pilot study was conducted with participants ranging in age from 12 to 19 for a preliminary evaluation.

CCS CONCEPTS

• **Human-centered computing** → **Scenario-based design**; • **Computing methodologies** → *Supervised learning by classification*; Cognitive science.

KEYWORDS

human-robot interaction, neural networks, social games

ACM Reference Format:
Baijun Xie and Chung Hyuk Park. 2023. "Can You Guess My Moves?" Playing Charades with a Humanoid Robot Employing Mutual Learning with Emotional Intelligence. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23 Companion), March 13–16, 2023, Stockholm, Sweden.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3568294.3580170

1 INTRODUCTION

Human emotion recognition is a vital aspect of human-robot interaction (HRI). An effective automatic emotion recognition system that can understand the underlying emotions of human behaviors is desired to establish a natural bidirectional interaction between users and robots. These human behaviors include multimodal inputs such as facial expressions, tone of voice, and body movements. Furthermore, it would be important to apply the emotion recognition system that can analyze users' affective states from multimodal social cues in a social HRI scenario.

Diverse deep learning methods have been widely used in recent years for emotion recognition. Previous studies recognized emotions through facial expressions [16, 34], acoustic features [10, 37], and texts [5, 9]. However, humans always express emotions with multimodal pathways and can achieve better performance via multimodal emotion recognition [39]. Humans interact with others primarily through speech while also coming up with body gestures to emphasize certain emotions in certain parts of speech [31]. Therefore, uni-modal emotion recognition can be incomplete and ambiguous under certain conditions, such as when human body language can convey different social cues than verbal communication. Previous researchers also investigated the usage of body movements to assess the affective states [1, 13, 26]. These studies demonstrated that body posture could enhance emotional expressions on the face and in speech. Furthermore, the robot can learn the gestures from humans as the robot behaves and benefit the HRI in terms of empathy [27], trust [29] and engagement [21, 23, 30]. Miura et al. [27] used functional magnetic resonance imaging to measure brain activity during the observation of emotional actions performed by humanoid robots, and the results showed that the robots with the ability to simulate human-like behaviors could elicit more empathy for the users. Reinhardt et al. [29] demonstrated the importance of movement as a non-verbal social cue that could be an indicator of trustworthiness. Ricks et al. [30] suggested that imitation interaction is effective for motivating and engaging users with autism spectrum disorder (ASD).

The capacity to highlight emotions using gestures is a vital social characteristic when communicating with people. Even in the absence of verbal communication, gestures may convey a variety of contexts in different social situations. Since it is a natural human trait to imitate others, imitation is a fundamental social process that helps children affiliate with others and develop empathy. The tendency for people to imitate the actions of the person they are communicating with is known as the "Chameleon Effect" [25]. Additionally, it was mentioned that imitation interaction effectively motivates and engages users with ASD [30].

In this sense, social games can be an effective method for teaching communication and social skills to children and adolescents

HRI '23 Companion, March 13–16, 2023, Stockholm, Sweden

ACM ISBN 978-1-4503-9970-8/23/03…$15.00
https://doi.org/10.1145/3568294.3580170

with social anxiety [28]. The social robot would play the role of a sympathetic playmate or mediator during HRI [22, 38]. SARs as interactive playmates have been widely used in autism therapy for children with ASD [11, 12, 33]. In autism therapy, social robots were usually involved in delivering engaging interactions to assist children with ASD in practicing social interaction skills [36]. More recent studies have explored the use of a serious game to reduce the stress and anxiety suffered by children with ASD [7, 8]. The results of the studies show the potential applications for reducing stress and anxiety for children with ASD. However, few studies have investigated the underlying emotional context during HRI to provide a more empathetic and engaging social robot for the user.

In this work, we propose a novel social game scenario with an interactive robot based on charades, a word-guessing game from body gestures. Our system employs a mutual learning paradigm that enables both the user and the robot to learn from the interaction. The robot will learn human behaviors through the interactive game, and the user can also benefit from creativity and social interaction. We also incorporate emotion recognition from human body movements to provide emotional intelligence for more personalized and social engagement. The goal of this study is to utilize the aspect of imitation to influence participants' physical, social, and mental behavior in a structured social setting where the participant and robot alternately engage in various gestural social interactions and imitation games. By utilizing our mutually-growing social robot and this mirroring process, we plan to provide an intervention for social engagement in autistic adolescents.

2 METHODOLOGY

In this section, we first present our emotion recognition model for detecting users' affective states during HRI. Detecting emotions will be used for personalizing the robot's behaviors and increasing engagement. Then, we demonstrated our proposed social game scenario with a mutual learning scheme.

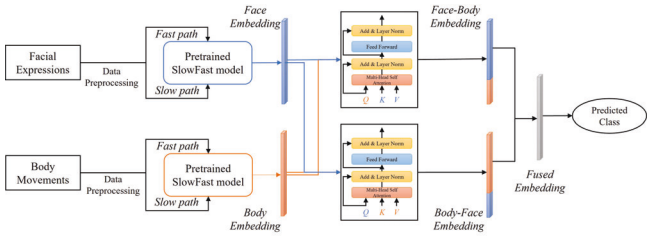

Figure 1: The overall system for bi-modal emotion recognition with embedding vectors (Q: query, K: key, and V: value).

As shown in Figure 1, the proposed model includes deep pre-trained convolutional neural networks (CNNs) and a model-level fusion architecture. The pre-trained CNNs are used as backbone models for fine-tuning and extracting the features from the face and upper body. In addition, a transformer-based neural network is introduced for the fusion strategy to achieve model-level fusion. The reason for employing facial inputs when the upper body contains the face part is that individuals may emphasize their emotions through their facial expressions, which can be a complement to the whole upper body's feature inputs.

2.0.1 Pre-trained Convolution Neutral Network. The pre-trained CNN model used in this study is the SlowFast network [14], which was designed for the task of video recognition. The SlowFast network structure comprises two pathways: a Slow pathway that operates at a low frame rate to learn the spatial semantic information of the video; and a Fast pathway that operates at a high frame rate to capture human motion at high temporal resolution. Both the Slow and Fast pathways are 3D ResNet [15], which operates 3D convolution and captures the information from sequential frames of image inputs. The Slow pathway is implemented with a large temporal stride τ, which extracts one frame out of τ frames. The typical value of τ is 16, allowing 2 frames per second for 30-fps videos. On the other hand, the Fast pathway uses a small temporal stride of τ/α, and a typical $\alpha = 8$ can process roughly 15 frames per second. Furthermore, to make the Fast pathway lightweight, it uses a significantly smaller channel size, with a ratio of β to the Slow pathway channel size, where the typical value of β is 1/8. The model used in this study is pre-trained on Kinetics-400 [24], a dataset containing 400 human action classes and more than 400 videos for each action category.

The bimodal face and body gesture (FABO) database [18] was used to fine-tune the emotion recognition model in this study. The FABO database captured the facial expressions and upper body movements via two cameras and annotated them with the affective states, including *happiness*, *surprise*, *anger*, *fear*, *sadness*, *disgust*, *boredom*, *puzzlement*, *uncertainty* and *anxiety*. Moreover, the annotations of the stages of the affective states were provided and separated into neutral, onset, apex, and offset states. For example, in one video, the subject began with a neutral state, acted with emotional gestures and expressions in the onset state, reached a steady level at the apex state, and ended with an offset state. Each video has two to four complete exhibition cycles. We extracted the onset-apex-offset cycle to represent the emotional states of the video, and a separate emotional state, "Neutral," is obtained based on the neutral state from all the videos.

2.0.2 Fusion Network. To achieve the fusion of the face and body modalities of inputs, we used the crossmodal Transformer to model the interaction of these two modalities. The crossmodal Transformer was first proposed by [35] to capture the correlated crossmodal signals for the task of natural language understanding. Following the definition of [35], for example, we denote the interaction modeling of introducing modality face, F, to modality body, B, as "$F \rightarrow B$". In the Transformer, the multi-head attention accepts query, key, and value as inputs and outputs the attention vector.

We denote the input features for Modalities F and B as $X_F \in \mathbb{R}^{T_F \times d_F}, X_B \in \mathbb{R}^{T_B \times d_B}$, where T and d are the time length and feature dimension. To achieve $F \rightarrow B$, we calculate the crossmodal attention via:

$$\begin{aligned} Y_{attention} &= Attention(X_B, X_F) \\ &= softmax(\frac{Q_B K_F^\top}{\sqrt{d_k}})V_F \\ &= softmax(\frac{X_B W_{Q_B} W_{K_F}^\top X_F^\top}{\sqrt{d_k}})X_F W_{V_F}. \end{aligned} \tag{1}$$

where queries are defined as $Q_B = X_B W_{Q_B}$, keys are defined as $K_B = X_B W_{K_B}$ and values are defined as $V_F = X_F W_{V_F}$. $W_{Q_B} \in \mathbb{R}^{d_B \times d_k}$, $W_{K_F} \in \mathbb{R}^{d_F \times d_k}$, $W_{V_F} \in \mathbb{R}^{d_F \times d_V}$ are the weights.

For both modalities, we used the same two SlowFast pre-trained models and fine-tuned them on the FABO database. Finally, to evaluate the robustness of our proposed method, we performed a 5-fold cross-validation and reported the results.

Table 1: The performance comparison of our proposed model with state-of-the-art methods. The reported results are unweighted. (Unit=%)

Propose method	Face	Upper Body	Bi-modal
Gunes and Piccardi [19]	35.2	73.1	82.7
Barros et al. [3]	72.7	57.8	91.3
Barros and Wermter [4]	87.3	74.8	93.65
Ilyas et al. [20]	90.42	79.27	94.41
Our proposed	92.08	90.2	94.79

As demonstrated in Table 1, our proposed method of using the fusion model for the bimodal results reached state-of-the-art methods, which reflected that combining the pre-trained models with the domain knowledge has the capabilities to capture salient features. Moreover, the results for the mono-modal upper body movements were considerably higher than the previous methods. This module will be used to measure users' emotions during the HRI.

2.1 Charades Game Design with Motion Imitation

2.1.1 Motion Imitation. To realize motion imitation in HRI, the robot platform needs to be presented, and an appropriate gesture mapping algorithm needs to be determined. We employed the Pepper robot platform manufactured by Softbank Robotics [17]. Pepper is a social humanoid robot designed for social interaction with humans through conversation, gestures, and touch screen capabilities.

For the joint control module, Pepper has NAOqi APIs that allow for the control of each joint angle. A reliable human pose estimation technique is necessary to detect the body of the user in order to give robots the capacity to mimic human movements. OpenPose is a real-time human pose estimation library that can detect keypoint skeletons on the human body, hand, and face in images or video [6]. The extracted 3D body pose keypoints will be used for motion retargeting, and the face keypoints will also be preserved.

Researchers in the past have developed several imitation algorithms for Pepper that can be used to mimic the user's movements [2, 40, 41]. Generally, the human posture skeleton keypoints were used to build the upper body link vectors of the Pepper robot, and the joint angles were then derived from the angles formed by those vectors. Further details can be found in the study by Zhang et al. [41]. The motion imitation module has been realized in real time for the teleoperation of Pepper.

2.1.2 Charades Game. Based on the charades game, we created an interactive social gaming scenario. At this point, five student researchers have provided the initial pre-defined words and related movement data. The Pepper robot can act on nine general words from the game of Charades, including the categories of sports, movies, and social behaviors. In addition to these general words, we also used emotion-based words from the gestural data in the FABO database.

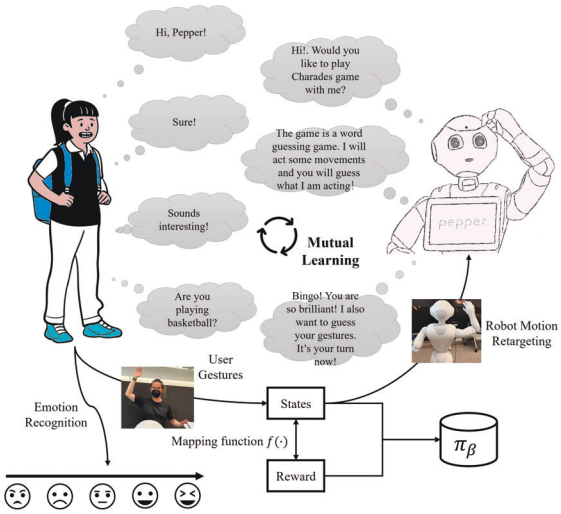

Figure 2: The interactive social game scenario that was designed based on the Charades game.

As shown in Figure 2, we intended to employ a mutual learning scheme to design the HRI scenario. During HRI, Pepper will first ask users whether they are willing to play the game with Pepper at the beginning of each session; if they accept, Pepper will begin performing the pre-defined charades words from the initially collected behavior database for users to guess. If users get the question right, Pepper will commend them; if they get it wrong, Pepper will nudge them to keep going and repeat gestures if they want. To provide variety to the interaction, Pepper can even act out the words' gestures performed by another person from the database. Then, once users agree to act out their charades, Pepper will attempt to guess what they are acting. Based on the developed motion retargeting module, Pepper will imitate users' movements in the meantime and solicit feedback from them regarding how well Pepper is mimicking. The user can benefit from the aspect of creativity when making their own charade movements with the Pepper robot imitation. This mirroring mechanism and mutually growing social robot are not just for fostering empathy but also for rehabilitation in adolescents with ASD, coinciding with the mutual learning scheme.

For the robot to understand the users' preferences, the emotion recognition module will continue to function during the encounter to recognize the underlying emotions of the users' behaviors. Based on the user's preferences, the learned policy can comfort them using the recognized emotions. For example, the policy can learn which category of charades words can elicit positive emotions from the users so that the robot can personalize the interaction in the next stage. Nevertheless, the primary goal of this study is to gather user data, which will be utilized to train the policy later.

3 PILOT STUDY

We conducted a pilot study to evaluate the robotic system with our proposed social game scenario (user study approved by the Institutional Review Board (GW IRB 111540)). The pilot study was conducted among four high school students (three males and one female). During each session, Pepper initiated the predefined charade words and let the participant guess; then, Pepper requested the participant to perform their charade words, which Pepper guessed in turn. The participants were invited to interact with Pepper to play the charades game as long as they wanted and were also informed that they were free to stop whenever they did not want to continue.

Table 2: The number of the charades words performed by Pepper and the participants.

	S1	S2	S3	S4
Pepper's Words	3	2	2	3
User's Words	3	1	4	4
Overall Interactions	6	3	6	7
Detected Positive Emotions	3	0	1	2

Table 2 demonstrates the number of charades words performed by Pepper and participants and the detected positive emotions by the multimodal emotion recognition model during the HRI. The proposed multimodal emotion recognition provides a measurement for evaluating the system. However, most of the detected emotions were neutral. Therefore, the model needs more personalized user behavior data to increase performance. Nevertheless, despite the limited number of users, it still shows the tendency that if Pepper acts with more words, the user will be more likely to act with their own words.

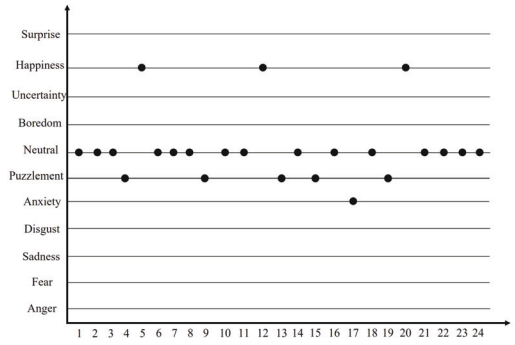

Figure 3: The change of emotions is detected by the proposed emotion recognition model. The x-axis is the time step, which is based on a window size of 30 seconds. Thus, the figure shows a 12-minute video that is segmented into 24 clips and fed into the model to detect emotion.

Figure 3 presents the change in emotions of one of the participants detected by the proposed emotion recognition model based on upper body movements. It can be seen that most of the detected emotions were neutral because the user hung his hands when standing. Three happy emotions were detected. However, five puzzlement emotions and one anxiety emotion were also detected.

We also used the Negative Attitude toward Robots Scale (NARS) [32] to measure the user's attitude toward and acceptance level of the robot. Three sub-scales make up the NARS: negative attitudes toward robot interaction scenarios, negative attitudes toward robot social influence, and negative attitudes toward robot interaction emotions. There are 17 items total, and each is evaluated from 1 (strongly disagree) to 5 (strongly agree). We calculated the overall NARS scale (M = 37.0, STD = 3.93), and we were primarily interested in the sub-scale of negative attitudes toward robot interaction emotions (M = 9.33, STD = 0.47), where the sub-scale includes all the reversed items from the scale. A lower score is desired for an emotional robot. We also investigated the relationship between the NARS scores and the overall count of interactions. However, there is a non-significant, very small negative relationship between the overall NARS scores and the number of interactions ($r = 0.041$, $p = 0.959$). Nevertheless, if the focus is on the sub-scale of negative attitudes toward robot interaction emotions, in that case, the results indicate a more significant negative relationship between the sub-scale scores and the number of interactions ($r = 0.802$, $p = 0.198$). Despite the limited sample size, we found that users with positive emotions toward the robot could play more during the interaction.

4 CONCLUSIONS

This study presented our novel social interaction scenarios for HRI. The scenarios were designed based on the charades game and combined with the emotion recognition module to recognize and collect more personalized data from users. The emotion recognition model could recognize the emotions based on the dataset with salient accuracy, even though it only depended on upper body movements. Furthermore, the imitation module for social games has been developed and can be run in real-time. The mutual learning framework enabled the robot to learn human behaviors through the proposed interactive game. We also created a custom behavior database to recognize contextual behaviors in social games. A pilot study was conducted to evaluate the proposed social game scenario, and the proposed emotion recognition model detected the user's emotions during the interaction. However, the emotion recognition model needs to be improved by fine-tuning users' personalized behavior data. Furthermore, the robotic agent can be personalized via actions that elicit positive emotions from users, which our proposed model can detect. We plan to conduct a further user study to evaluate the impact of the socio-emotional interventions for autistic individuals with this novel interactive game we proposed.

As a next step, we will conduct a user study to evaluate the designed social gaming scenario to alleviate participants' anxiety. The user study for future work will be conducted among teenagers between 13 and 19 years old. The participants will be recruited from high school or as freshmen at the university.

ACKNOWLEDGMENTS

The authors would like to appreciate the National Science Foundation for supporting this research (NSF 1846658).

REFERENCES

[1] Ferdous Ahmed, ASM Hossain Bari, and Marina L Gavrilova. 2019. Emotion recognition from body movement. *IEEE Access* 8 (2019), 11761–11781.

[2] Archana Balmik, Mrityunjay Jha, and Anup Nandy. 2021. NAO Robot Teleoperation with Human Motion Recognition. *Arabian Journal for Science and Engineering* (2021), 1–10.

[3] Pablo Barros, Doreen Jirak, Cornelius Weber, and Stefan Wermter. 2015. Multimodal emotional state recognition using sequence-dependent deep hierarchical features. *Neural Networks* 72 (2015), 140–151.

[4] Pablo Barros and Stefan Wermter. 2016. Developing crossmodal expression recognition based on a deep neural model. *Adaptive behavior* 24, 5 (2016), 373–396.

[5] Erdenebileg Batbaatar, Meijing Li, and Keun Ho Ryu. 2019. Semantic-emotion neural network for emotion recognition from text. *IEEE Access* 7 (2019), 111866–111878.

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7291–7299.

[7] Stéphanie Carlier, Sara Van der Paelt, Femke Ongenae, Femke De Backere, and Filip De Turck. 2019. Using a serious game to reduce stress and anxiety in children with autism spectrum disorder. In *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*. 452–461.

[8] Stéphanie Carlier, Sara Van der Paelt, Femke Ongenae, Femke De Backere, and Filip De Turck. 2020. Empowering children with ASD and their parents: Design of a serious game for anxiety and stress reduction. *Sensors* 20, 4 (2020), 966.

[9] Ankush Chatterjee, Umang Gupta, Manoj Kumar Chinnakotla, Radhakrishnan Srikanth, Michel Galley, and Puneet Agrawal. 2019. Understanding emotions in text using deep learning and big data. *Computers in Human Behavior* 93 (2019), 309–317.

[10] Mingyi Chen, Xuanji He, Jing Yang, and Han Zhang. 2018. 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Processing Letters* 25, 10 (2018), 1440–1444.

[11] Andreia P Costa, Georges Steffgen, Francisco Rodríguez Lera, Aida Nazarikhorram, and Pouyan Ziafati. 2017. Socially assistive robots for teaching emotional abilities to children with autism spectrum disorder. In *3rd Workshop on Child-Robot Interaction at HRI*.

[12] Kerstin Dautenhahn. 2003. Roles and functions of robots in human society: implications from research in autism therapy. *Robotica* 21, 4 (2003), 443–452.

[13] Beatrice de Gelder, AW De Borst, and R Watson. 2015. The perception of emotion in body expressions. *Wiley Interdisciplinary Reviews: Cognitive Science* 6, 2 (2015), 149–158.

[14] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6202–6211.

[15] Christoph Feichtenhofer, Axel Pinz, and Richard Wildes. 2016. Spatiotemporal residual networks for video action recognition. In *Advances in Neural Information Processing Systems (NIPS)*. 3468–3476.

[16] Panagiotis Giannopoulos, Isidoros Perikos, and Ioannis Hatzilygeroudis. 2018. Deep learning approaches for facial emotion recognition: A case study on FER-2013. In *Advances in hybridization of intelligent methods*. Springer, 1–16.

[17] SoftBank Robotics Group. 2022. *Pepper the humanoid and programmable robot: SoftBank Robotics*. Retrieved December 7, 2022 from https://us.softbankrobotics.com/pepper

[18] Hatice Gunes and Massimo Piccardi. 2006. A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In *18th International conference on pattern recognition (ICPR'06)*, Vol. 1. IEEE, 1148–1153.

[19] Hatice Gunes and Massimo Piccardi. 2008. Automatic temporal segment detection and affect recognition from face and body display. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39, 1 (2008), 64–84.

[20] Chaudhary Muhammad Aqdus Ilyas, Rita Nunes, Kamal Nasrollahi, Matthias Rehm, and Thomas B Moeslund. 2021. Deep Emotion Recognition through Upper Body Movements and Facial Expression.. In *VISIGRAPP (5: VISAPP)*. 669–679.

[21] Hifza Javed, Rachael Burns, Myounghoon Jeon, Ayanna M Howard, and Chung Hyuk Park. 2019. A robotic framework to facilitate sensory experiences for children with autism spectrum disorder: A preliminary study. *ACM Transactions on Human-Robot Interaction (THRI)* 9, 1 (2019), 1–26.

[22] Hifza Javed and Chung Hyuk Park. 2019. Interactions with an empathetic agent: regulating emotions and improving engagement in autism. *IEEE robotics & automation magazine* 26, 2 (2019), 40–48.

[23] Hifza Javed and Chung Hyuk Park. 2022. Promoting Social Engagement with a Multi-Role Dancing Robot for In-Home Autism Care. *Frontiers in Robotics and AI* (2022), 161.

[24] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017).

[25] Jessica L Lakin, Valerie E Jefferis, Clara Michelle Cheng, and Tanya L Chartrand. 2003. The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. *Journal of nonverbal behavior* 27, 3 (2003), 145–162.

[26] Katie Lang, Marcela Marin Dapelo, Mizanur Khondoker, Robin Morris, Simon Surguladze, Janet Treasure, and Kate Tchanturia. 2015. Exploring emotion recognition in adults and adolescents with anorexia nervosa using a body motion paradigm. *European Eating Disorders Review* 23, 4 (2015), 262–268.

[27] Naoki Miura, Motoaki Sugiura, Makoto Takahashi, Tomohisa Moridaira, Atsushi Miyamoto, Yoshihiro Kuroki, and Ryuta Kawashima. 2008. An advantage of bipedal humanoid robot on the empathy generation: A neuroimaging study. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2465–2470.

[28] Samira Rasouli, Garima Gupta, Elizabeth Nilsen, and Kerstin Dautenhahn. 2022. Potential applications of social robots in robot-assisted interventions for social anxiety. *International Journal of Social Robotics* (2022), 1–32.

[29] Jakob Reinhardt, Aaron Pereira, Dario Beckert, and Klaus Bengler. 2017. Dominance and movement cues of robot motion: A user study on trust and predictability. In *2017 IEEE international conference on systems, man, and cybernetics (SMC)*. IEEE, 1493–1498.

[30] Daniel J Ricks and Mark B Colton. 2010. Trends and considerations in robot-assisted autism therapy. In *2010 IEEE international conference on robotics and automation*. IEEE, 4354–4359.

[31] Nicu Sebe, Ira Cohen, and Thomas S Huang. 2005. Multimodal emotion recognition. In *Handbook of pattern recognition and computer vision*. World Scientific, 387–409.

[32] Dag Sverre Syrdal, Kerstin Dautenhahn, Kheng Lee Koay, and Michael L Walters. 2009. The negative attitudes towards robots scale and reactions to robot behaviour in a live human-robot interaction study. *Adaptive and emergent behaviour and complex systems* (2009).

[33] Adriana Tapus, Mataric Maja, and Brian Scassellatti. 2007. The grand challenges in socially assistive robotics. *IEEE Robotics and Automation Magazine* 14, 1 (2007), N–A.

[34] Paweł Tarnowski, Marcin Kołodziej, Andrzej Majkowski, and Remigiusz J Rak. 2017. Emotion recognition using facial expressions. *Procedia Computer Science* 108 (2017), 1175–1184.

[35] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, Vol. 2019. NIH Public Access, 6558.

[36] Iain Werry, Kerstin Dautenhahn, and William Harwin. 2001. Investigating a robot as a therapy partner for children with autism. *Procs AAATE 2001,* (2001).

[37] Baijun Xie, Jonathan C Kim, and Chung Hyuk Park. 2020. Musical emotion recognition with spectral feature extraction based on a sinusoidal model with model-based and deep-learning approaches. *Applied Sciences* 10, 3 (2020), 902.

[38] Baijun Xie and Chung Hyuk Park. 2021. Empathetic robot with transformer-based dialogue agent. In *2021 18th International Conference on Ubiquitous Robots (UR)*. IEEE, 290–295.

[39] Baijun Xie, Mariia Sidulova, and Chung Hyuk Park. 2021. Robust multimodal emotion recognition from conversation with transformer-based crossmodality fusion. *Sensors* 21, 14 (2021), 4913.

[40] Unai Zabala, Igor Rodriguez, José María Martínez-Otzeta, and Elena Lazkano. 2022. Modeling and evaluating beat gestures for social robots. *Multimedia Tools and Applications* 81, 3 (2022), 3421–3438.

[41] Zhijun Zhang, Yaru Niu, Ziyi Yan, and Shuyang Lin. 2018. Real-time whole-body imitation by humanoid robots and task-oriented teleoperation using an analytical mapping method and quantitative evaluation. *Applied Sciences* 8, 10 (2018), 2005.