Personal and Ubiquitous Computing
https://doi.org/10.1007/s00779-022-01684-y

ORIGINAL PAPER

What is it like to be a bot? Variable perspective embodied telepresence for crowdsourcing robot movements

Michael Suguitan[1] · Guy Hoffman[1]

Received: 14 July 2021 / Accepted: 5 April 2022

Abstract
Movement and embodiment are communicative affordances central to social robotics, but designing embodied movements for robots often requires extensive knowledge of both robotics and movement theory. More accessible methods such as learning from demonstration often rely on physical access to the robot which is usually limited to research settings. Machine learning (ML) algorithms can complement hand-crafted or learned movements by generating new behaviors, but this requires large and diverse training datasets, which are hard to come by. In this work, we propose an embodied telepresence system for remotely crowdsourcing emotive robot movement samples that can serve as ML training data. Remote users control the robot through the internet using the motion sensors in their smartphones and view the movement either from a first-person or a third-person perspective. We evaluated the system in an online study where users created emotive movements for the robot and rated their experience. We then utilized the user-crafted movements as inputs to a neural network to generate new movements. We found that users strongly preferred the third-person perspective and that the ML-generated movements are largely comparable to the user-crafted movements. This work supports the usability of telepresence robots as a movement crowdsourcing platform.

Keywords Human-robot interaction · Affective computing · Telepresence · Neural networks · Crowdsourcing

1 Introduction

Duffy et al. define a “social robot” as “a physical entity embodied in a complex, dynamic, and social environment sufficiently empowered to behave in a manner conducive to its own goals and those of its community” [9]. Social robots can communicate through their embodiment and movements, which serve to not only achieve utilitarian functions but also to convey affective states [15]. Movement is an important nonverbal communication modality that differentiates robots from graphics- or voice-based agents. However, designing robot movements is often a costly process that requires expertise in robotics and movement theory. Accessible methods such as learning from demonstration (LfD) enable lay-users to provide movement samples by either physically manipulating the robot or controlling its degrees-of-freedom (DoFs) [4, 34]. In some cases, larger sample libraries can be elicited using crowdsourcing methods [5, 21, 22]. Movement libraries, whether hand-generated, crowdsourced, or learned, can be further expanded with generative models that analyze existing samples and synthesize new realistic movements [8, 28, 32, 43] (Fig. 1). For example, deep neural networks can learn important data features given a sufficient diversity of samples, thus relaxing the need for expert knowledge in movement generation [31]. As a result, human-robot interaction (HRI) researchers have begun applying neural networks for generating robot movements [23, 37, 44], but these approaches are limited by the availability of data.

Restrictions on in-person experiments due to the COVID-19 pandemic forced HRI researchers to shift towards remote technologies, such as simulators or telepresence robots, and this shift could prove beneficial for robot movement generation. Researchers have also used these remote technologies to conduct online evaluations and crowdsource data. Services such as Amazon Mechanical Turk and Prolific have enabled the collection of data from a diverse user base. Paired with telepresence

✉ Michael Suguitan
mjs679@cornell.edu

Guy Hoffman
hoffman@cornell.edu

[1] Cornell University, Ithaca, USA

Published online: 04 May 2022

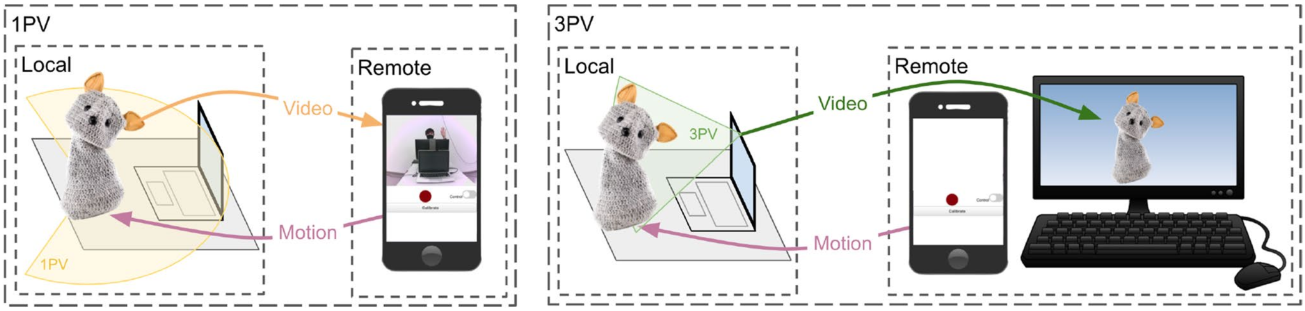


Fig. 1 Roboticists can complement their initially small set of handcrafted movements by crowdsourcing new samples from users. Machine learning techniques can then further expand the available movements by generating new samples. This work focuses on the crowdsourcing and generation aspects

platforms, crowdsourcing could also enable the collection of user-crafted demonstrations for robots. A machine learning model could then use the collected data to generate new samples and further expand the robot's behavior library.

In this work, we present a system for remotely crowdsourcing emotive robot movements through a telepresence robot. The robot is controlled with a smartphone, a widely accessible device that enables a direct mapping from the user's body to the robot using the phone's built-in motion sensors. We compared two alternate viewpoints for the interface: a through-the-robot first-person view (1PV) seen on the phone, and a whole-body third-person view (3PV) seen on an external monitor (Fig. 2). We performed an evaluation where users controlled the robot and recorded emotive movements to collect a diverse user-crafted data set. To validate the usability of the collected data set for ML movement generation, we trained a neural network to generate new movements, and deployed a survey to subjectively compare the user-crafted and generated movements. Our contributions are:

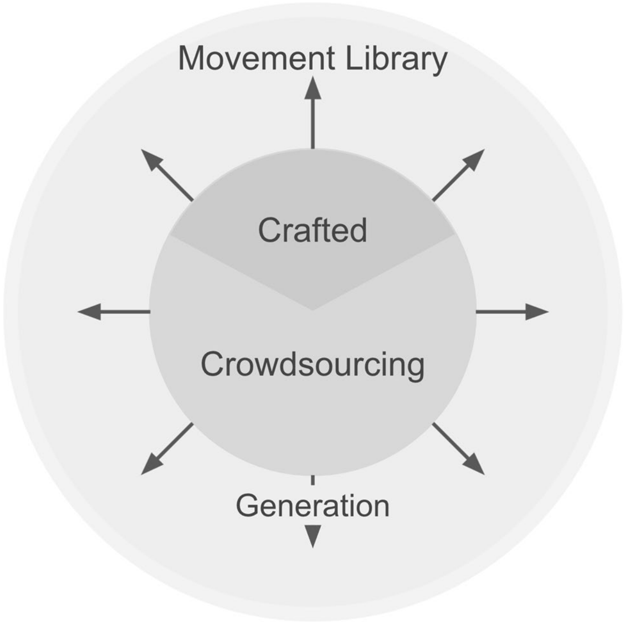


Fig. 2 In the first-person view (1PV, left) the camera feed is transmitted from the local robot to the remote phone. In the third-person view (3PV, right) the local computer camera feed capturing the robot is transmitted to the desktop. In both cases, the remote phone's motion data is transmitted to the local computer to control the robot's motors

- An accessible system for remotely motion controlling a robot in either the first- or third-person, requiring no specialized hardware.
- An evaluation of the system as an embodied telepresence platform. We conducted a remote study for users to control the robot, create emotive movements, and rate their experience using the platform comparing the first- and third-person views.
- An evaluation of the quality of the user-crafted movements as a data set for ML generation, first by using a generative neural network to synthesize new movement samples, then by deploying a survey to compare the user-crafted and generated samples.

2 Related work

In this section we review related works in affective telepresence, teleoperation control methods, and crowdsourcing for robot learning.

2.1 Affective telepresence

The physicality of objects can promote nonverbal and ludic interactions beyond the affordances of visual or auditory communication modalities. Strong and Gaver's Feather, Scent, and Shaker were minimally expressive home objects for technologically mediated sociality between remote users [36]. More specifically within robotics, Goldberg's early telepresence robots emphasized playful interactions, such as tending to a garden or uncovering treasures in a sandbox by remotely controlling a robot through the internet [11]. Sirkin and Ju found that augmenting a screen-based telepresence robot with motion improved the sense of presence on both ends [33]. Tanaka et al. compared video, avatar, and robot communications and found that the presence and movements of a robot improved the conversation partner's sense of social presence [40]. The teddy bear Huggable robot enabled remote users to control its gaze and appendages through a web interface [35]. Gomez et al. used the Haru robot for transmitting "robomojis," emojis that are embodied by the robot's motion, animations, and sounds [12]. The MeBot telepresence robot features controllable appendages in addition to a screen displaying the remote user [1]. Similarly, Tsoi et al. created a phone application to turn the Anki Vector robot into a telepresence platform controlled with game-like touchscreen joysticks; this work was a direct response to the sudden isolation of children due to COVID-19 safety restrictions [42]. While these embodied platforms afford an additional dimension of engagement beyond virtual agents, most use button- or joystick-centric controllers that abstract remote users away from their own bodies as a communicative medium. Employing the user's own embodied movement is possible through motion control.

2.1.1 Motion control

Rather than use text inputs or game controllers as proxies for controlling robots, proprioceptive motion controls afford a more direct translation between the embodiments of the user and robot, enhancing the sense of self-location and agency [18]. Ainasoja et al. compared motion- and touch-based smartphone interfaces for controlling a Beam telepresence robot, and found that users preferred a hybrid motion-touch interface (motion for left-right steering, touch for forward-reverse) [3]. Jonggil et al. compared touch and motion controls for a mobile camera robot, and found that motion controls improved the user's sense of presence, synchronicity, and understanding of the remote space [17]. In a more affective application, Sakashita et al. used a virtual reality system with head and arm tracking to remotely embody and puppeteer robots [30]. Many of these robots were utilitarian in design and function, require specialized hardware, and the user perspectives were constrained to first-person views.

2.1.2 Viewpoint control

In traditional video chat applications, the remote user's view is controllable only by their interaction partner. Müller et al. created a panoramic stitching application to enable remote users to freely adjust their view by panning their phone around the environment, and found that this significantly improved measures of spatial and social presence and slightly improved copresence [24]. Tang et al. extended this work by replacing the panoramic stitching with a 360° camera [41]. They recommended improvement to collocation, such as indicators to dictate gaze direction or ways to convey remote gestures. Young et al. combined the panoramic stitching and 360° camera into a single evaluation while also adding the user's hand into the shared view as a gesture indicator, and found that both implementations increased spatial presence and copresence [45]. Free choice between first- and third-person is a common interface setting in video games, and several works have shown that first-person perspectives increase immersion and the sense of body ownership while third-person offers heightened spatial awareness [7, 10, 13, 20]. To our knowledge, viewpoint effects on experiential factors of telerobotics operation have not been thoroughly explored.

2.2 Crowdsourcing demonstrations for robots

Robotic systems can implement LfD systems that enable lay-users to provide high-fidelity data for machine learning models. However, collecting demonstrations is still time-consuming and often constrained by physical proximity to a robot. Mandlekar et al. created a system for remotely crowdsourcing grasping task demonstrations for simulated and physical robot arms, and found that more data improves model performance [22]. Among various input devices ranging from mice to virtual reality controllers, they found smartphones to be the best compromise of accessibility and functionality. The primary performance metric was grasp success, with completion time as a secondary measure. Timing is an important feature for affective expression, specifically the arousal dimension on the circumplex model of emotions [29]. Rakita et al. found that while users could adapt to a teleoperated robot's physical slowness, latency between the user's movement and the robot executing the motion reduced performance, further emphasizing the importance of timing [27].

2.3 Research questions

There are several gaps in existing works. Prior works focused primarily on the usability of different control methods, but were either constrained to first-person perspectives or designed for utilitarian, nonaffective functions. Alternatively, we are interested in fixing the control input and instead varying the viewpoints. Although prior works measured subjective experiential responses from the users as both operators and interactors with the robot, many did not focus on the affective quality of the robot's movements. Additionally, there are few prior works in enabling remote crowdsourcing of robot movement demonstrations. We address these gaps by designing a robot telepresence system with accessible motion controls and variable viewpoints. We perform user evaluations to assess the subjective usability of the system for creating emotive robot movements. We then use the movements to train a neural network to generate new movement samples, and perform another evaluation to compare the user-crafted and generated movements. This work probes the following research questions:

- Would affective telepresence be better achieved with a first- or third-person perspective?
- Are crowdsourcing movement demonstrations and generative neural networks viable methods for expanding a robot's behavior library?

3 Technical implementation

In this section we detail the technical implementation of the system, including the robot and user interfaces.

3.1 Robot

We used the Blossom robot, an open-source social robot (Fig. 2, left) [38]. Blossom's internal mechanisms consist of a head platform suspended from a tower structure that rotates about its base platform. Blossom features four DoFs: yaw, pitch, roll, and vertical translation, though we disable vertical translation to simplify the control interface. The robot achieves motion with four actuators: tower motors 1, 2, and 3 control the front, left, and right sides of the head, respectively, and a motor in the base rotates the tower left and right. The robot's head can pitch up and down and roll left and right ±45° and can yaw 300° left and right about its base. Although the robot's DoFs are limited compared to more complex embodiments, it features a large range of motion (RoM) and head movements alone can convey complex affective information [2]. For 1PV, we embedded a small USB camera inside the robot's head, in front of one of its ears. The camera has a wide-angle lens (21 mm equivalent, 95° diagonal angle of view) to maximize the viewing range.

3.2 User interfaces

To bolster the system's accessibility, we built the application as a mobile browser experience instead of creating a standalone application. This enabled us to iterate quickly and access a rich library of functionality through APIs while obviating the need for external downloads on the user's device. We created two interfaces to accommodate the two viewpoints (Fig. 2, right): a mobile interface showing 1PV from the camera in the robot's head, and a desktop interface showing 3PV from the host computer's webcam. Users access both interfaces from a public URL. 1PV for the mobile interface acts as a "window" *through* the robot; 3PV for the desktop interface acts as a "mirror" *at* the robot.

3.2.1 Mobile interface

The mobile interface consists of a video feed showing 1PV and a simple layout of buttons for controlling the robot (Fig. 2, center). The layout was inspired by existing controlling and recording interfaces, such as camera applications and voice recorders. Control of the robot is toggled with a slider switch. Users can record and save movements with a large microphone-style recording button. The robot can be reoriented using a calibration button; this resets the robot's yaw orientation relative to the phone's current compass heading, setting it to face towards the external camera. If the user rotates to the endpoints of the base RoM, indicator arrows appear on the interface to direct the user back towards the center.

3.2.2 Desktop interface

The desktop interface consists of a video screen showing 3PV (Fig. 2, right). The mobile interface still controls the robot, but 1PV is hidden to force users to look *at* the robot instead of *through* the robot. For the evaluation (described later in Section 4.1), the interface also features a YouTube video player, controls for displaying a video from a given URL, and a Qualtrics survey at the bottom of the page.

3.3 Back end

3.3.1 Communication

The robot is connected to the host computer, which also serves the interfaces. We use `ngrok` to enable communication across the internet from the user to the host computer and robot[1]. We open two `ngrok` tunnels: one for accessing the user interfaces, and another for transmitting the phone orientation data to motion control the robot.

3.3.2 Motion control

Kinematic models of the phone and robot translate the phone's orientation into the angular poses of the robot's head (Fig. 3). The mobile interface uses the

[1] https://ngrok.com/

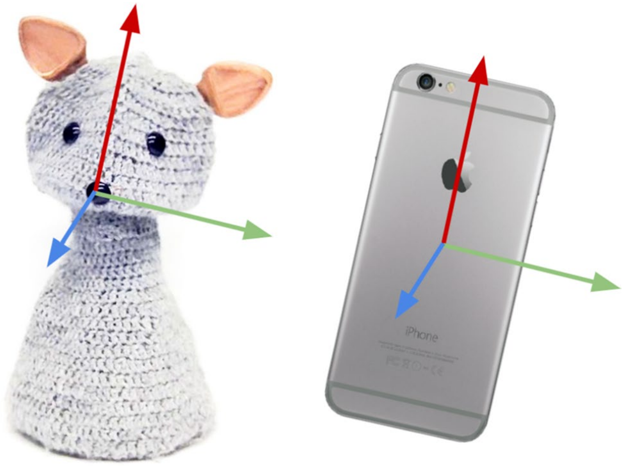


Fig. 3 The alignment of the robot and phone reference frames when controlling in 1PV. In 3PV, the motion is mirrored to accommodate the perspective of looking straight at the robot (e.g., motion towards the phone's left moves the robot to its right)

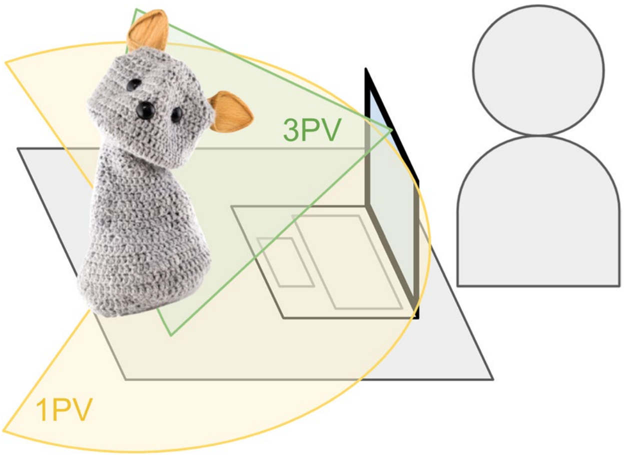


Fig. 4 Evaluation setup showing the fields of view of 1PV (yellow) and 3PV (green). The evaluation proctor (right) acts as a focal point when controlling the robot in 1PV

`DeviceOrientation` API to report motion events[2]. The phone's inertial measurement unit (IMU) records its pose as Tait-Bryan angles about the phone's reference frame. In 1PV, the phone and robot axes are aligned as if the phone's camera were looking through the robot's eyes. When switching from 1PV to 3PV, the motion is mirrored horizontally to accommodate the front-facing view of the robot, as if the user were facing a physical mirror. In 3PV, yawing or rolling the phone to the left from the user's perspective moves the robot to its right, and vice versa. Assuming a stable connection, motion data is transferred at a rate of approximately 10 Hz.

3.3.3 Video

For the video streams, we use `WebRTC`, the standard for online audiovisual communication[3]. `WebRTC` manages the handshaking for broadcasting the local video stream to remote viewers.

4 Experiments

In this section we detail the evaluations of the interface and the crowdsourced dataset.

4.1 Interface evaluation

We measured the usability of the system and compared 1PV and 3PV through an online user evaluation for a movement creation task (Figs. 4, 5). We recruited participants through the Prolific online survey platform[4]. Apart from limiting enrollment to users within the United States (for latency concerns), we did not record any demographic information. We first instructed the user to navigate to the interfaces on both their phone and desktop. The user connected to the robot and tested the controller by looking around the environment in 1PV, then in 3PV. Only one viewpoint (1PV on the phone, 3PV on the desktop) is visible at a time. Because of the importance of timing for the task, we measured the latency between when the orientation data packet is sent from the user's phone and when it is received by the robot's host computer. This latency is only "one-way" and is exacerbated by the video latency, so the user will experience a longer delay from their perspective. Latency below 100 ms is very good and around 1,000 ms (1 s) is serviceable, but exceeding 2,000 ms (2 s) noticeably degrades usability. If the user's latency exceeded the 2-s threshold, we would end the study prematurely and compensate the user proportionally to their time spent.

For the main movement creation task, we had the users record examples of emotive gestures. We prompted the users with short videos, between 5 and 10 s in length, of cartoon characters (either SpongeBob, Pikachu, or Homer Simpson) displaying either happiness, sadness, or anger. We then had the users control the robot to express the emotion from the video and record the movement. We urged users to not simply mimic the motion of the characters, but rather to move the robot as if it were conveying the overall emotion from the scene. Users could rehearse and re-record the movements until they were satisfied, but could not redo the movement once they moved on to the next video. We introduced two

[2] https://www.w3.org/TR/orientation-event/

[3] https://www.w3.org/TR/webrtc/

[4] https://www.prolific.co/

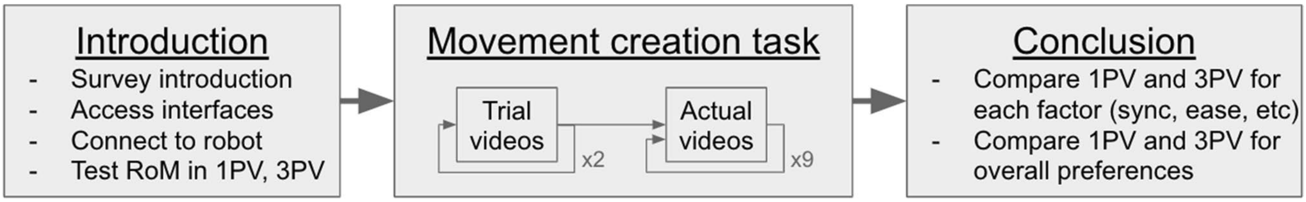


Fig. 5 Interface evaluation flow. Users first access the interfaces and test the robot's motion. In the main movement creation task, users watch videos of cartoon characters emoting, then create movements for the robot corresponding to the conveyed emotions (happy, sad, or angry). The evaluation concludes with a comparative assessment of the perspectives for user experience factors and overall preferences

Table 1 Interface evaluation survey questions, displayed after every video and again at the end of the survey to compare 1PV and 3PV. *Note: scales for mental and physical tiredness are reversed from how they were displayed in the evaluation (1 = not tiring, 7 = tiring) to better match the other factors*

Question	1 (low rating)	7 (high rating)
How synchronized with the robot did you feel?	Unsynchronized	Synchronized
How much did you feel present in the remote location?	Separate	Present
How easy was controlling the robot?	Difficult	Easy
How enjoyable was controlling the robot?	Not enjoyable	Enjoyable
How engaging was controlling the robot?	Not engaging	Engaging
How mentally tiring was controlling the robot?	Tiring	Not tiring
How physically tiring was controlling the robot?	Tiring	Not tiring
How do you feel about the quality of the movement you created?	Low quality	High quality

trial videos to acclimate the user to the task, followed by nine actual videos (three emotions for each of three different characters). To account for learning effects, we randomized the video orders and perspectives so that each would be equally represented (e.g., four 1PV and five 3PV, or vice versa). We measured the latency during recording for post-analysis of its effect on the user experience.

We used surveys throughout and after the evaluation to collect user-reported metrics. After each video, we asked for subjective 7-point Likert scale responses to measure experiential factors (Table 1). After all of the videos, we again asked for Likert scale responses for each factor, but asked for comparative responses for both 1PV and 3PV. We also asked for overall preferences between the perspectives and included a free response field for any additional feedback. Due to the limited expressiveness of the robot platform, we expected differences across the different emotion classes (e.g., sadness will be more homogeneous but easier to convey than anger). We preregistered hypotheses regarding the experiential factors[5]:

H1.1 1PV will increase the sense of synchronization with the robot due to a heightened sense of embodiment.

H1.2 1PV will increase the sense of presence in the remote location due to higher immersion.

H1.3 3PV will be easier to use due to heightened spatial awareness.

H1.4 1PV will be more enjoyable due to being a unique experience.

H1.5 1PV will be more engaging due to having to move around in one's physical space.

H1.6 1PV will be more mentally tiring due to having to embody a remote system with latency.

H1.7 1PV will be more physically tiring due to having to move one's whole body to maintain a view of the video.

H1.8 3PV will increase the self-reported quality of created movements due to being able to see the full robot.

We enrolled 30 participants through the Prolific platform and offered US $10 as compensation. We prescreened by participants with access to both a mobile device and desktop. In the interest of minimizing latency, we restricted enrollment to participants living in the United States. We proctored the evaluation through an audio-only Zoom call and took approximately 30 min to complete: 10 min for the introduction and 20 min for creating the movements. We occasionally encountered incompatibilities with certain Android devices, often stemming from access permissions for the orientation sensor. In cases where we were unable to troubleshoot the problem, we ended the study prematurely and compensated the participants proportionally to their time spent; this led us to eventually prescreen to users with Apple devices. We did not have to reject any participants on the basis of high latency.

[5] Preregistration link: https://aspredicted.org/pu8p3.pdf

4.2 Movement kinematic evaluation

We calculated kinematic features for each movement: length, speed, and range. Length is the overall duration of the movement, measured in seconds. Because there may have been delays between when the user pressed the record button and actually began or stopped moving, we trimmed the “whitespace” of no motion at the beginning and end of each movement. Speed is the angular velocity of the motors, measured in radians per second. Range is the wideness of the motion in each DoF, measured in radians. We averaged the speed and range across all DoFs for the entire movement. We preregistered hypotheses for the movement features:

H2.1 1PV will yield longer movements due to having to move around in one’s physical space.

H2.2 3PV will yield faster movements due to requiring less full-body motions.

H2.3 3PV will yield wider, more exaggerated movements due to requiring less full-body motions.

4.3 Dataset evaluation

To appraise the validity and usability of the system as a data collection platform, we used the user-crafted movements to train a neural network to generate new movements. The network architecture consists of a convolutional variational autoencoder (VAE) with an additional emotion classifier (Fig. 6) [19]. The VAE encodes the movement samples into a compressed lower-dimension latent embedding space (Fig. 6, left), then decodes these embeddings back into a reconstruction of the original samples (Fig. 6, bottom path). The classifier operates on the embeddings and separates the latent space by emotions (happy, sad, or angry) (Fig. 6, top path). We split the collected dataset by perspective (1PV and 3PV) and trained the network with identical parameters on both subsets. The technical results can be objectively evaluated in terms of the network training metrics, quality of the movement reconstructions, and separability of the emotion classes in the latent embedding space.

We compared the user-crafted and generated movements in a survey to appraise realism, emotiveness, and emotional legibility. We recorded the robot performing the movements from an external perspective similar to 3PV in the first evaluation, and thus used only movements created or generated with the 3PV dataset. We randomly selected subsets of user-crafted movements from a held-out test set and generated movements from the neural network. To avoid using several similar or static movements, we further manually curated the movements to four diverse and representative examples for each condition, resulting in a set of 24 movements (3 emotions × [User, Generated] × 4 examples). Users watched the movements and gave ratings for realism, emotiveness, and which emotion was conveyed (Table 2). We preregistered hypotheses for the movement comparison:

H3.1 The generated movements will be as realistic as the user-crafted movements.

H3.2 The generated movements will be as emotive as the user-crafted movements.

H3.3 The generated movements will be recognized with the same accuracy as the user-crafted movements.

Table 2 Movement comparison survey questions for comparing the user-crafted and generated movements

1 (low rating)	7 (high rating)
Fake	Natural
Emotionless	Emotional
Please select the emotion that best describes the robot’s movement	Happy, Sad, or Angry

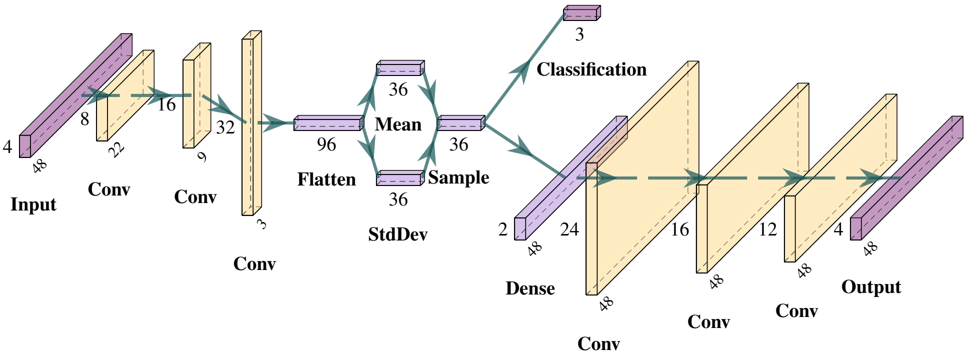


Fig. 6 Neural network architecture for generating movements. The user-crafted movements (4.8 s at 10 Hz with four DoFs → 48 × 4) are used as inputs and encoded into a 36D embedding space (left). The embeddings are both decoded to reconstruct the original input (bottom path) and classified into one of the three emotion classes (happy, sad, or angry) (top path)

5 Results

5.1 Interface evaluation results

We used two-sided *t*-tests to test **H1** from the end-survey Likert scale responses, and found that many results were significant in the *opposite* direction of our hypotheses favoring 1PV (Fig. 7). **H1.3** and **H1.8** were supported in the hypothesized direction. **H1.1**, **H1.4**, and **H1.5** were supported *opposite* of the hypothesized direction. We found overwhelming preference for 3PV, with significant results in synchronization, ease, enjoyment, engagement, and quality. Even increased presence, which we assumed would be decisively in favor of 1PV, is not supported. We also tested the hypotheses within each emotion class using the responses after every video, and only found slight support for sadness being more physically tiring in 1PV (Table 3). Interestingly, the within-emotion scores do not correlate with the comparative end-survey scores. The overall preferences are also favorable toward 3PV (Fig. 8).

We compared the end-survey scores against the average latencies for each user and for each perspective to analyze latency's effect on the experience (Fig. 9). As suggested by the low r^2 values, we found no correlation between latency and any factors, suggesting that latency did not noticeably affect the user experience.

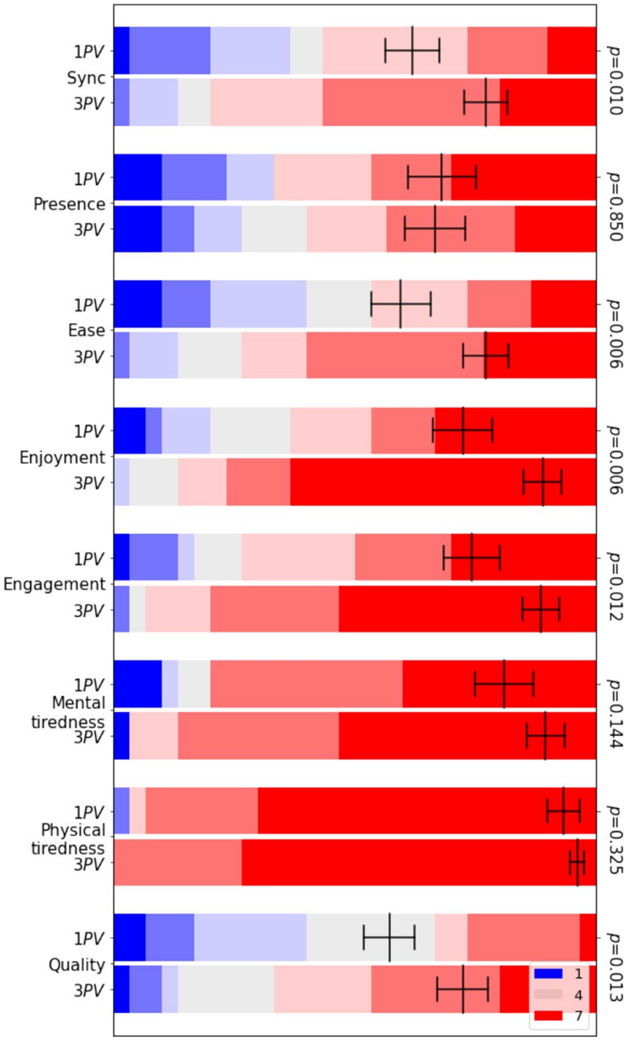


Fig. 7 Likert scale responses from the interface evaluation end-survey questions. Color indicates level: blue = 1 (low), gray = 4 (neutral), red = 7 (high). Width indicates proportion of responses for a given level. Black bars indicate means and standard deviations. *p*-values of **H1** tested with two-sided *t*-tests are displayed on the right, and the means indicate preferences for 3PV in all factors except presence andtiredness. Note: as in Table 2, the scales for mental and physical tiredness are reversed from what was displayed in the survey

5.2 Movement kinematic evaluation results

We computed the average kinematic features for each user and for each perspective, and used two-sided *t*-tests to test **H2** (Fig. 10). We found support for 3PV yielding faster and wider movements (**H2.2-3**), but no support for 1PV yielding longer movements.

5.3 Dataset evaluation results

The interface evaluation yielded approximately 135 movement samples from each perspective. We prepared the data by chunking the 4-DoF 10 Hz movements into samples of

Table 3 *p*-values of **H1** tested with two-sided *t*-tests within each emotion, calculated from the average of the scores after each video. Slight support is suggested only for sadness being more physically tiring in 1PV

Factor	Happy	Sad	Anger
Sync (1PV>3PV)	0.706	0.995	0.681
Presence (1PV>3PV)	0.365	0.667	0.911
Ease (3PV>1PV)	0.428	0.665	0.430
Enjoyment (1PV>3PV)	0.750	0.637	0.558
Engagement (1PV>3PV)	0.881	0.382	0.630
Mental tired (3PV>1PV)	0.567	0.960	0.619
Physical tired (3PV>1PV)	0.938	0.088	0.718
Quality (3PV>1PV)	0.908	0.744	0.609

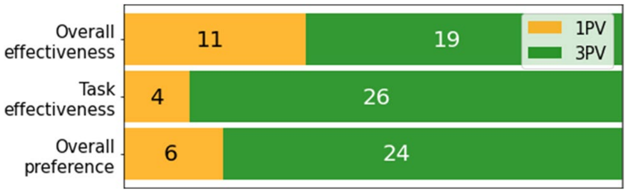


Fig. 8 Overall preferences reported at the end of the evaluation, showing strong preferences for 3PV

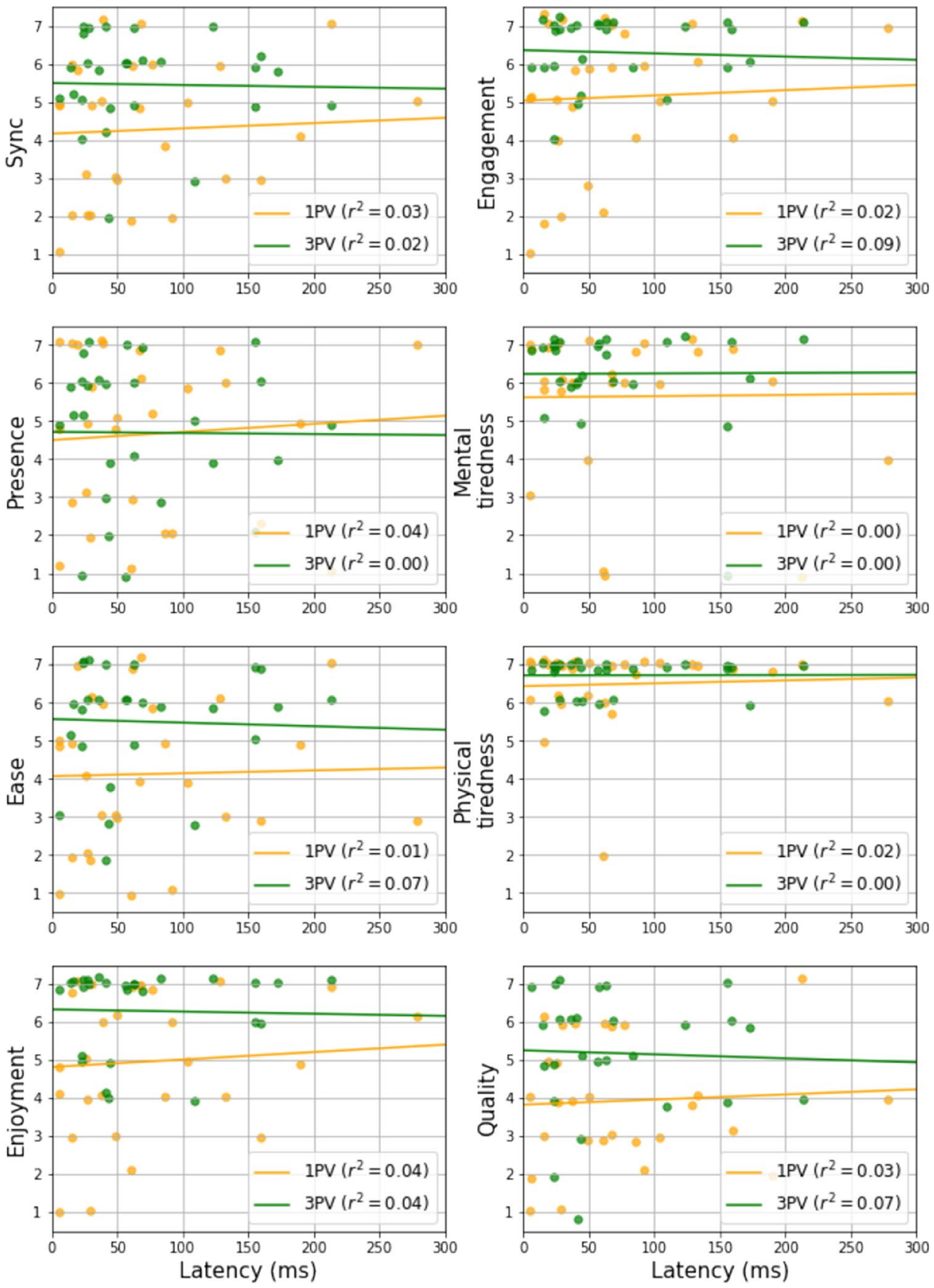


Fig. 9 Interface evaluation scores versus latency for each user for each perspective. The horizontal axes are truncated to 300 ms (maximum 900 ms) and vertical jitter is applied for legibility. The low r^2 values suggest no correlation between latency and any of the experiential factors

4.8 s with a sliding window of 0.3 s, resulting in 48 × 4 data samples. We then performed an 80-20 train-test split and augmented the training data by mirroring (flipping left-right), shearing (nudging the timing of DoFs relative to each other), shifting the center (adding small variation to the left-right direction that the robot is looking), and decoupling the left and right tower motors (preventing these DoFs from copying each other), yielding over 150, 000 training samples for each perspective. We tuned the neural network architecture and parameters until satisfactory results could be achieved on the datasets from both perspectives.

We empirically found that an embedding size of 36 was the lowest before noticeably degrading reconstruction performance. The encoder convolutions have a stride of 2 to progressively increase the effective receptive field. We trained the network for 10 epochs with a learning rate of 2×10^{-3} and a batch size of 32. We used Leaky ReLU activations ($\alpha = 0.01$), batch normalization [16], and 10% dropout

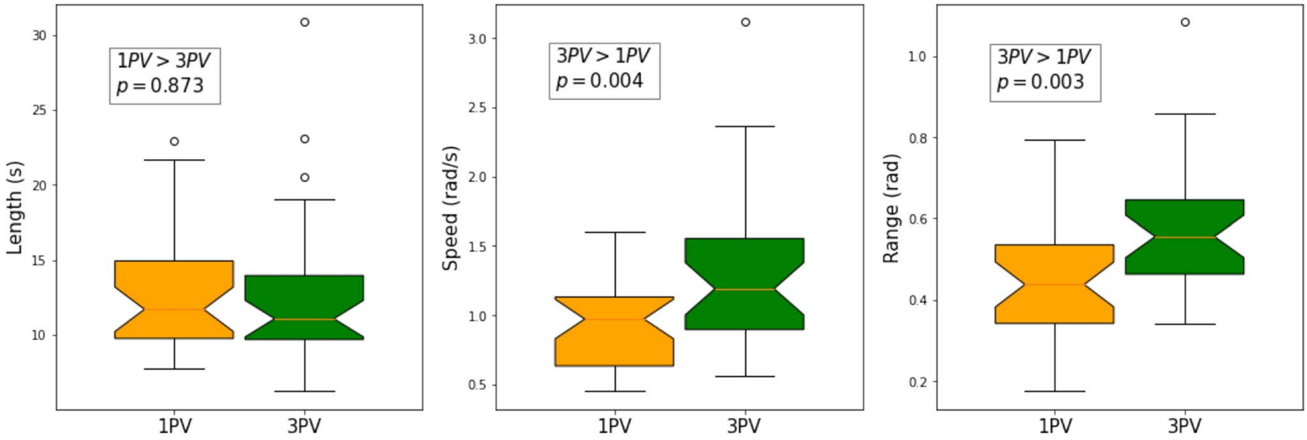


Fig. 10 Comparison of kinematic features between 1PV and 3PV, testing **H2** with two-sided *t*-tests. Movement length did not significantly vary between perspectives, but 3PV yielded faster and wider movements compared to 1PV

after the convolutional and dense layers, as well as a mixup parameter of 0.2 [46]. For the reconstruction loss, we used mean absolute error for the front (tower 1) and base DoFs, mean squared error for the side (towers 2 and 3) DoFs, and weighed the errors as 5, 7, and 10 for the front, side, and base DoFs, respectively. For the classification loss, we used categorical cross entropy on the softmax output of the classifier. For the overall loss, we applied weights of 5 and 7 for the reconstruction and classification losses, respectively, and implemented a $\beta = 0.1$ weight for the VAE's Kullback-Leibler divergence [14].

5.3.1 Network training results

We trained the networks on both datasets with varying dataset sizes as an ablation study (Fig. 11). We found that the 3PV dataset required less tuning to achieve better results. There is a noticeable improvement for the overall loss compared to using only 10% of the dataset, but only marginal improvement compared to using 50%. While it appears that smaller training datasets do not dramatically impact classification accuracy, the testing dataset sizes were also decreased; the high classification accuracies with smaller datasets are actually "overfit" and thus less generalizable to unseen samples.

5.3.2 Movement reconstruction results

We compared movement reconstruction accuracy with varying dataset sizes (Fig. 12). Reconstruction fidelity increases with more data, most noticeably in the base motion. The network captures the overall trajectories but has difficulty achieving the same level of exaggeration and reconstructing granular motions, such as low-amplitude high-frequency jitter.

5.3.3 Embedding separability results

We used t-SNE to further compress the 36D embeddings into visualizable 2D representations (Fig. 13). As corroborated by the classification accuracies, the emotion clusters are more separable in the 3PV dataset than the 1PV dataset. This suggests that the 3PV movements are more diverse and will yield more emotionally legible generated movements.

5.3.4 Movement generation results

To generate new movements, we first randomly sampled about the embedding distributions of each emotion (e.g., for a new happy movement, we sampled a 36D embedding about the mean and standard deviation of the happy embeddings), then passed these embeddings through the VAE decoder to generate full 48×4 movements. Upon inspection, the generated movements look comparable to the user-crafted movements (Fig. 14). We performed an objective comparison for calculable kinematic features (Fig. 15). Equivalence tests with bounds of $\pm 20\%$ of the range for a given feature show similarities in pitch (the amount of upward and downward tilt, measured in radians) for all emotions, acceleration for sadness, and range for happiness and sadness. We subjectively evaluated the comparability through a user survey.

5.3.5 Movement comparison survey results

We deployed the movement comparison survey on Prolific, offered US $2 in compensation for approximately 10 min of work, and received 100 responses. Each user watched and rated 15 random movements out of the total set of 24 movements. We averaged each user's responses for each emotion, source, and measure, then rounded to the nearest integer on the Likert scale (e.g., a given user's responses for realism

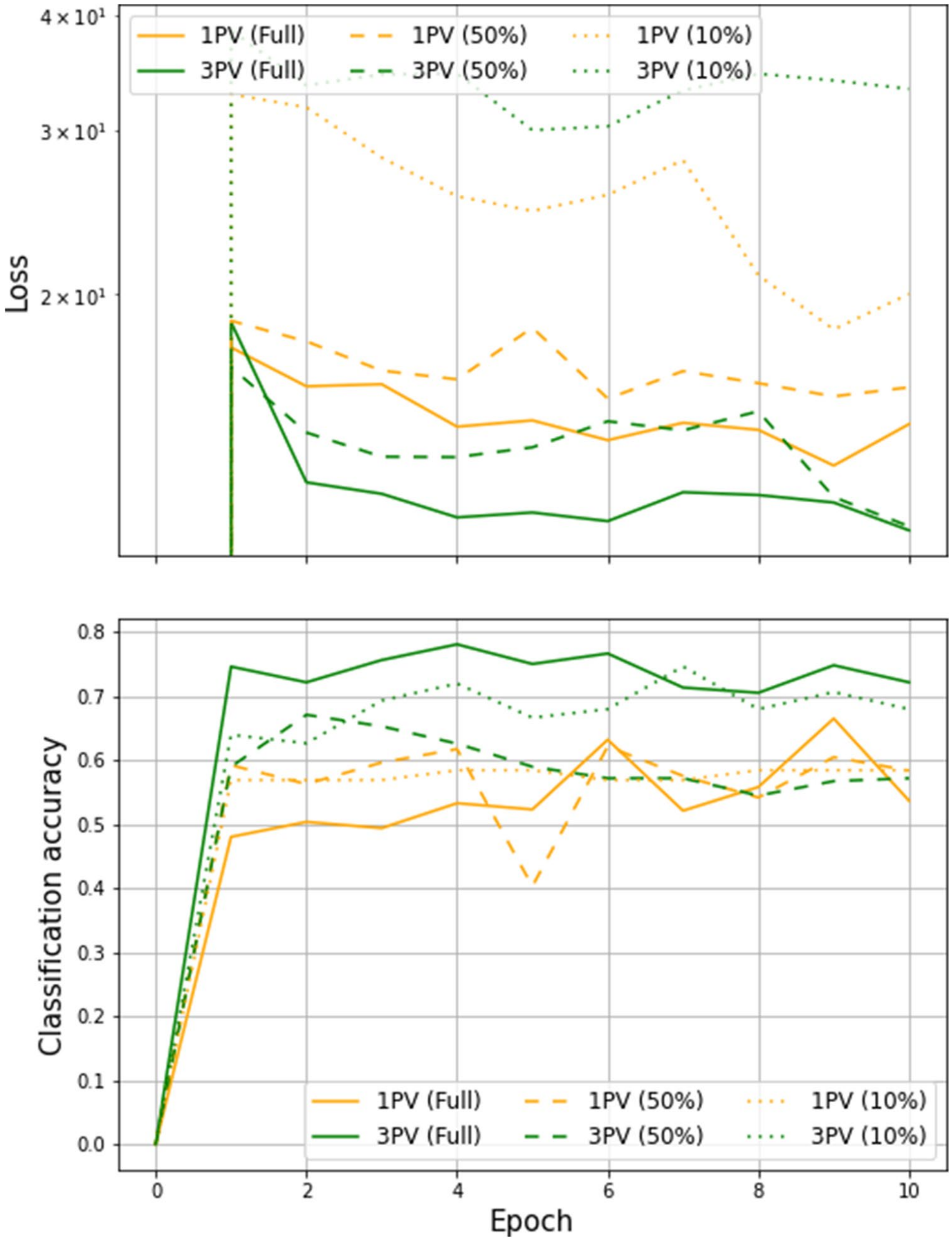


Fig. 11 Network training results on the test sets. Color indicates perspective, line style indicates data size. Using more data generally lowers the overall loss (top), but only slightly improves classification accuracy (bottom). The small improvement indicates that the network “overfits” to the smaller test set when using less data

for all happy user-crafted videos they saw are averaged and rounded into a single Likert score, which represents one data point used in the top left bar of Fig. 16). On the unrounded per-user averages, we used equivalence tests (two one-sided *t*-tests) with an equivalence bound of 0.6 (1/10th of the 7-point Likert scale) to test **H3**. The results show that the generated movements are comparable to the user-crafted movements in many measures, supporting **H3.1-2**, except for user-crafted happy movements being more emotive and angry movements being more realistic. In the context of the prior objective kinematic comparison, these results suggest that pitch is a particularly important feature in conveying affect.

We compared the recognition rates between the actual and interpreted emotions (Fig. 17). The recognition accuracies are well above chance (33%) for both the user-crafted and generated movements. Looking at the row-wise results, happiness and sadness are recognized with high accuracies, supporting **H3.3**, though generated happy movements are more ambiguous. Anger has low recognition rates in both sources, and the column-wise responses indicate that users selected anger much less frequently than the other emotions.

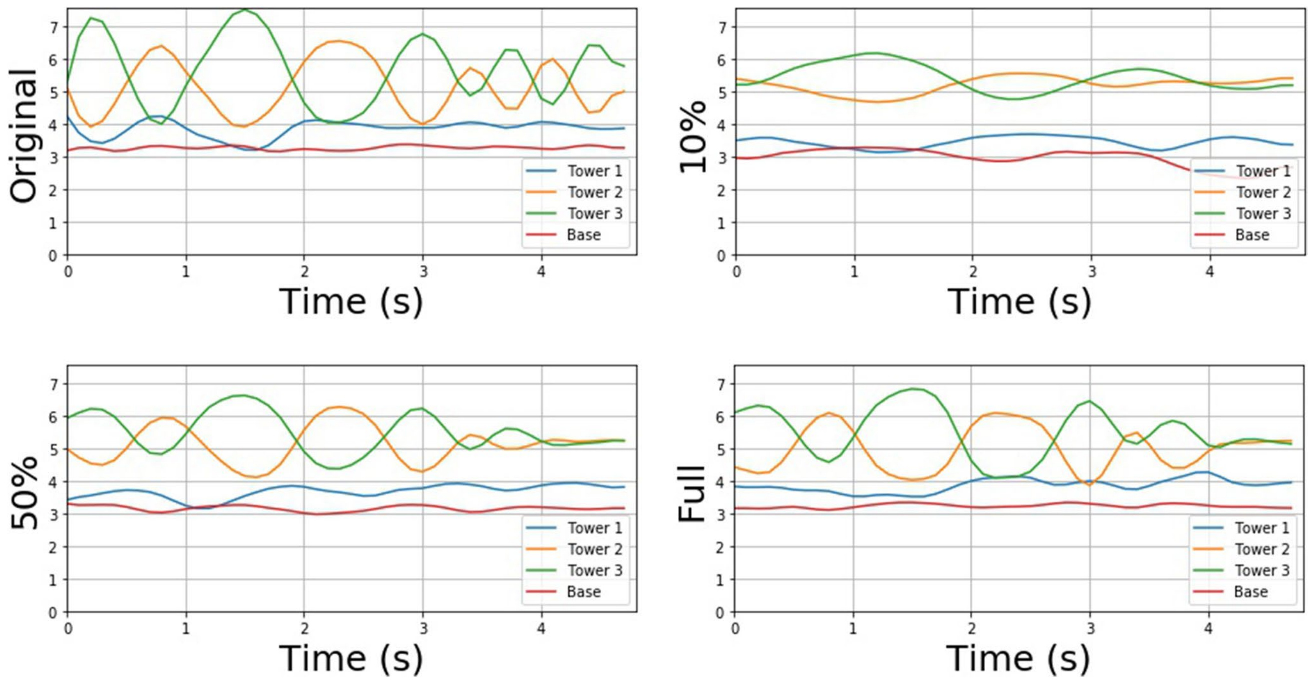
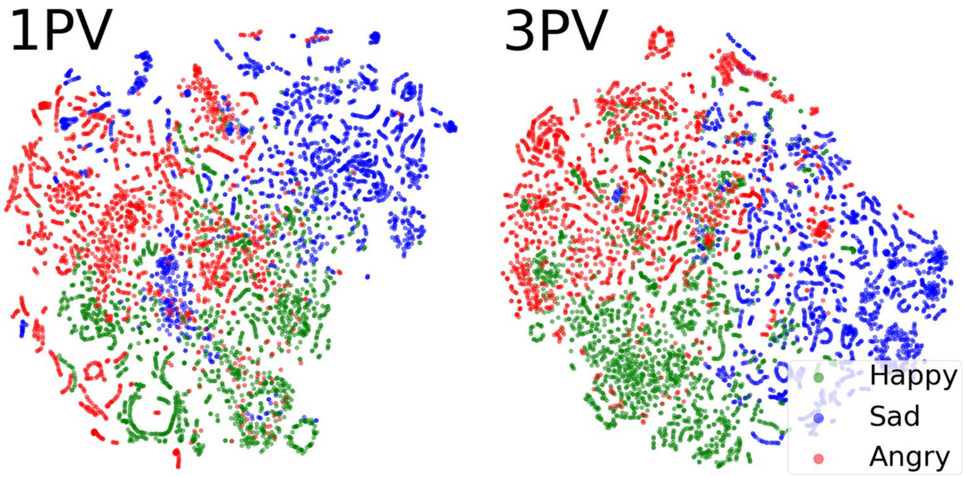


Fig. 12 Movement reconstructions with varying 3PV dataset sizes. Reconstruction fidelity is proportional to dataset size

Fig. 13 Embedding space visualization using t-SNE for 1PV (left) and 3PV (right), color-coded by emotion (happy = green, sad = blue, angry = red). 3PV is more separable, suggesting more diversity and legibility



6 Discussion

The interface evaluation revealed strong preferences for 3PV, suggesting that an external perspective may be more useful for conveying affect remotely. The dataset evaluations showed that the user-crafted movements are usable as inputs to the neural network for generating new movements. The movement comparison survey supported movement generation as a valid approach for expanding a robot's behavior library.

Feedback to the interface evaluation was largely positive, with many participants commenting on the uniqueness and enjoyability of the experience. Several participants also commented on the robot's design, remarking on its cuteness and the fun factor in controlling the robot remotely. The robot's aesthetic appeal may explain the strong preferences for being able to watch it move in 3PV.

Latency can explain the lower than expected synchronization and presence measures in 1PV. Compared to viewing

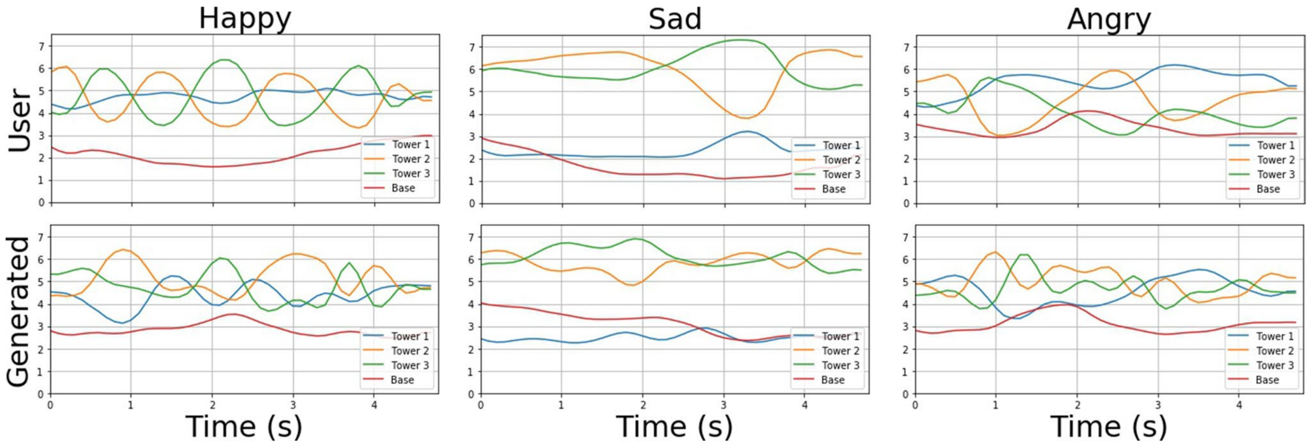


Fig. 14 Sample trajectories of user-crafted (top) and network-generated (bottom) movements from the 3PV dataset. The generated movements retain the characteristics of the original user-crafted movements

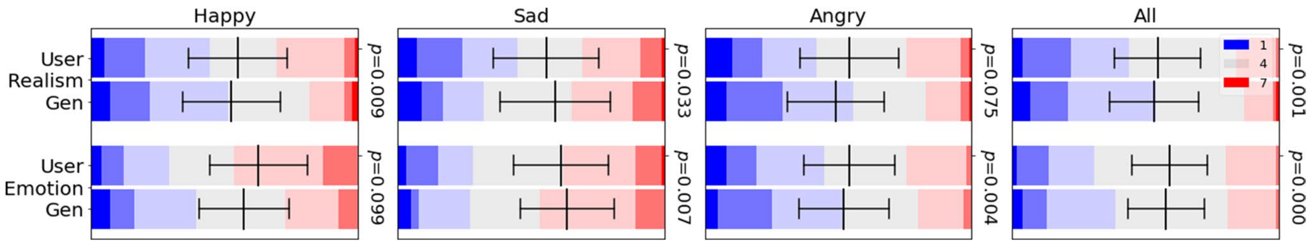


Fig. 15 Kinematic comparison between user-crafted and generated movements, with equivalence test scores (bounds set to ±20% of the range for a given feature) annotated on each graph. The tests show similarities primarily in pitch (the amount of upward and downward tilt, measured in radians)

the external robot in 3PV, 1PV may heighten the expectation of synchrony between motion and the video updating. Latency lands 1PV in a temporal uncanny valley, exacerbating the delay and negatively affecting the experience.

Latency can also explain the slower, smaller movements in 1PV. Although we did not view the users during the evaluation, it is reasonable to posit that 1PV employs more of the user's body as they must turn their head to maintain a view of the video. In contrast, control in 3PV requires only hand and arm movements, which enables users to create faster and wider movements.

The neural network training results support performance increasing with more data, though our dataset is still magnitudes smaller than publicly available datasets for common modalities such as images or text. There are relatively few works in generative affective robot movements that generalize across different robot platforms and machine learning methods. Establishing standardized comparisons for generative movement algorithms is important for future research to build upon prior works; the GENEA Project is a recent development that aims to address this issue by providing common datasets for benchmarking [21].

The subjective comparisons of the user-crafted and generated movements show that they are largely comparable, but also indicate limitations of the robot's embodiment, particularly when emoting anger. The low survey responses for anger and user feedback regarding the robot's limitations, specifically its lack of appendages and difficulty in tracking finer motions, indicate that more DoFs are necessary for delineating subtleties in affect. Interestingly, the network classifier can outperform the human classifications (>70% compared to >60%), suggesting that the network learns latent features that are not legible from the movement videos.

6.1 Limitations and future work

6.1.1 Latency

Latency is the largest bottleneck in the system, but is the hardest to mitigate. Although the latency measurements for the trip from the user's phone to the robot's host computer could reach as low as 10 ms, we cannot accurately measure the return latency between when the robot moves and when the video

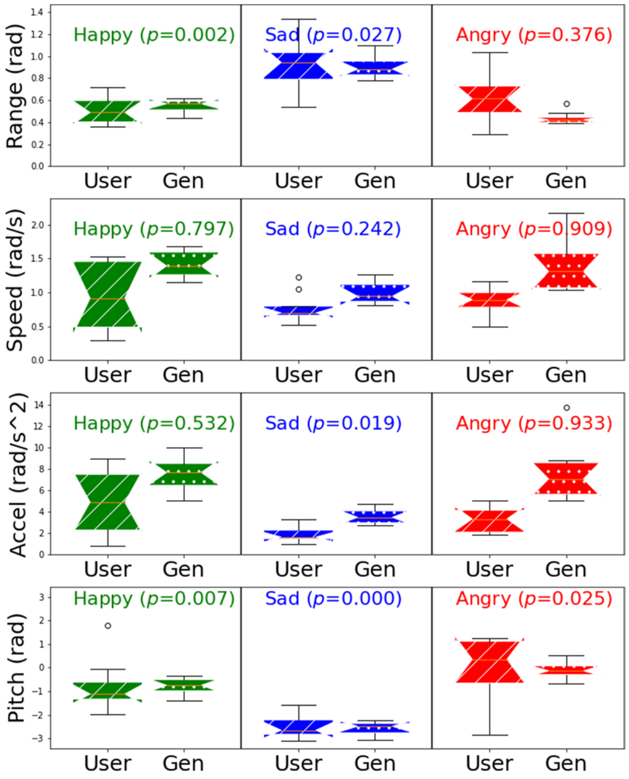


Fig. 16 Likert scale responses from the movement comparison survey. As in Fig. 7, color indicates level, width indicates proportion of responses for a given level, and black bars indicate means and standard deviations. For each user, the scores for each emotion (happy, sad, or angry) and source (user-crafted or generated) are calculated and rounded to the nearest integer (e.g., a given user's responses for realism for all happy user-crafted videos they saw are averaged and rounded into a single Likert score, which represents one data point used in the top left bar). p-values of **H3** tested with equivalence tests (two one-sided t-tests, equivalence bound of 0.6) are displayed on the right sides. The two sources are largely comparable, except for user-crafted happy movements being more emotive and angry movements being more realistic

updates on the user's device. WebRTC benchmarks measured round-trip times from 400 ms on a cellular network down to below 100 ms on a dedicated university connection [39]. By contrast, virtual reality systems are expected to perform with latency below 50 ms, and ideally below 20 ms [26]. Future technical work could involve optimizing the underlying technologies to minimize the latency, and perhaps even freely adjust latency as a controlled variable to investigate its effects on the user experience.

6.1.2 Embodiment

While the simplicity of the robot's design enabled novice users to quickly learn the control scheme, it also limited its expressive capabilities to three DoFs. Several users noted feeling that many of their movements were very similar and expressed wanting arms to convey strong emotions, particularly anger; additionally, the robot's "cuteness" may have limited its expressive range. The robot's vertical translation and ear DoFs were removed to simplify the interface, but these motions may be significantly important for affording more expressiveness.

6.1.3 Remote evaluation paradigm

Due to the social distancing restrictions that were in place at the time of this work, we designed the interface evaluation to focus solely on the experience of the remote participant. This neglects studying the experience of a local participant interacting with the robot, and how a remote participant would use the system accordingly. A two-sided scenario may reveal favorable situations for 1PV, such as tasks requiring joint attention or communication in a real-time environment.

6.2 Design implications

6.2.1 Research

Through this work, we gathered a dataset of affective movements from novice users, who provided usable samples after a short trial to acclimate to the system. The results of the interface evaluation suggest that 3PV is more enjoyable and useful for the movement generation task; future affective telepresence systems may benefit from this external perspective. The comparison survey results showed that these movements are still legible to other users, and support crowdsourcing and generation as viable methods for expanding a robot's given behavior library. Other researchers can adopt this accessible crowdsourcing approach for their own systems. For example, video-based pose trackers (e.g., OpenPose, VideoPose3D [6, 25]) can translate human motions into movements for humanoid robots [44], emancipating these systems from specialized motion capture environments. In the vein of RoboTurk [22], the remote control scheme could be adapted to source demonstrations for other LfD tasks such as locomotion or manipulation. Such open-access systems will require enforceable review policies to ensure the quality and usability of the samples, such as the two-survey approach with independent populations that we undertook in this work.

6.2.2 Fictional scenario

We imagine robots as a communicative medium that affords a transmission of one's physicality, adding an extra dimension beyond voice- or video-based mediums (Fig. 18). In one example scenario[6], two family members in separate locations

[6] This assumes that such social robotic systems are adopted on a similar scale as modern computing devices, either through commercial viability or open-sourcing.

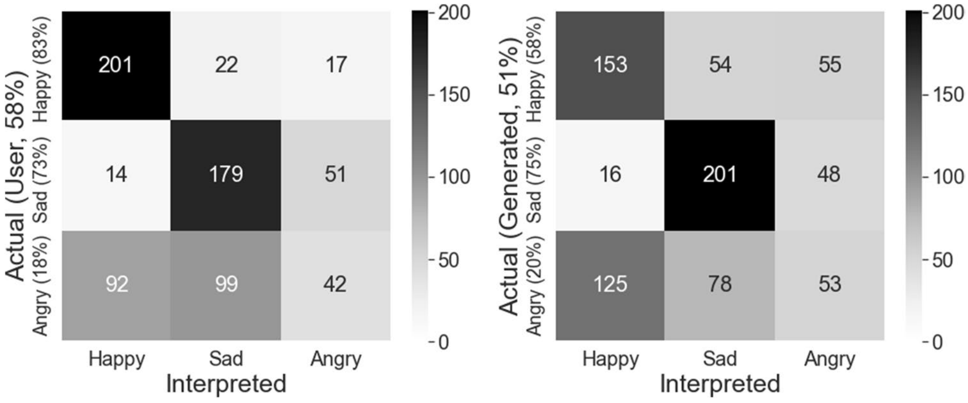


Fig. 17 Confusion matrices for user-crafted (left) and generated movements (right) using 3PV. Overall and within-emotion accuracies accompany the vertical labels. Happiness and sadness are largely correctly matched in both sources, but anger is rarely chosen

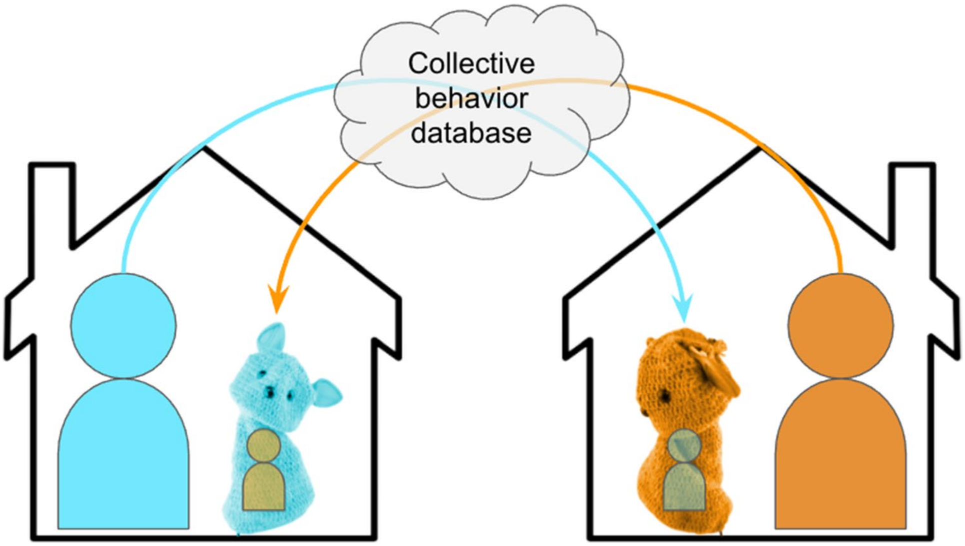


Fig. 18 Scenario depicting remote communication through pairs of robots in separate locations. Each user remotely controls their conversation partner's robot and can record behaviors, which are stored in a personal repository on each robot and in a collective database. Coupled with behavior generation algorithms, these behaviors imbue the robot with personalities that either reflect a specific user or represent the robot as a unique individual character

communicate through their conversation partner's respective robot, transmitting their voice, movement, and, optionally, their face through screens implemented on the robots. The remote users can record their movements and save them to their personal repository on their communication partner's robot. These movements are tied to a unique individual user, but are also added to a collective database of all user-crafted movements. The backend movement generation algorithm trains on both the individual and collective samples. With the individual samples, the robot learns to act as a proxy of a specific user by generating movements in their personal idiosyncratic style. With the collective samples, the robot learns to act as a unique individual character. While movement is seemingly more innocuous than incendiary imagery or text, future work may involve safeguarding against such adversarial content.

7 Conclusion

We presented a variable perspective telepresence system for motion controlling a social robot and crowdsourcing affective movement samples. The system uses a smartphone as an accessible motion-based input device. Users controlled the robot from one of two perspectives: either embodying the robot from a first-person perspective through a camera in the robot's head, or a third-person perspective with an external camera looking at the whole body of the robot. To crowdsource robot movements and assess the experiential quality of the system, we performed an evaluation where lay-users created emotive movement samples for the robot. The subjective responses showed strong preferences for the third-person perspective in self-reported measures of synchronization, ease, enjoyment, engagement, and quality of the created movements. The third-person perspective also yielded movements that were faster and wider than those created in the first-person. To evaluate the usefulness of the collected dataset, we used the user-crafted movements as inputs to a neural network to generate new movements. Through a second user survey, we found that the user-crafted and generated movements were largely comparable. This work supports the use of affective telepresence systems as crowdsourcing platforms for robot demonstrations, and hopefully inspires creative approaches for conducting remote human-robot interaction research.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Adalgeirsson S, Breazeal C (2010) MeBot: A robotic platform for socially embodied telepresence. In: 2010 5th ACM/IEEE International conference on human-robot interaction (HRI). pp 15–22. https://doi.org/10.1109/HRI.2010.5453272
2. Adams A, Mahmoud M, Baltrušaitis T, Robinson P (2015) Decoupling facial expressions and head motions in complex emotions. In: 2015 International conference on affective computing and intelligent interaction (ACII). pp 274–280. https://doi.org/10.1109/ACII.2015.7344583
3. Ainasoja AE, Pertuz S, Kämäräinen J-K (2019) Smartphone teleoperation for self-balancing telepresence robots. In: VISIGRAPP (4: VISAPP). pp 561–568 https://doi.org/10.5220/0007406405610568
4. Argall BD, Chernova S, Veloso M, Browning B (2009) A survey of robot learning from demonstration. Robotics and Autonomous Systems 57(5):469–483. https://doi.org/10.1016/j.robot.2008.10.024
5. Breazeal C, DePalma N, Orkin J, Chernova S, Jung M (2013) Crowdsourcing human-robot interaction: new methods and system evaluation in a public environment. J Hum-Robot Interact. 2(1):82–111. https://doi.org/10.5898/JHRI.2.1.Breazeal
6. Cao Z, Hidalgo G, Simon T, Wei S-E, Sheikh Y (2019) OpenPose: realtime multi-person 2D pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence 43(1):172–186. https://doi.org/10.1109/TPAMI.2019.2929257
7. Denisova A, Cairns P (2015) First person vs. third person perspective in digital games: do player preferences affect immersion?. In: Proceedings of the 33rd annual acm conference on human factors in computing systems. (CHI '15). association for computing machinery, New York, NY, USA, 145–148. https://doi.org/10.1145/2702123.2702256
8. Desai R, Anderson F, Matejka J, Coros S, McCann J, Fitzmaurice G, Grossman T (2019) Geppetto: enabling semantic design of expressive robot behaviors. In: Proceedings of the 2019 CHI conference on human factors in computing systems. (CHI '19). ACM, New York, NY, USA, Article 369, 14 pages. https://doi.org/10.1145/3290605.3300599
9. Duffy BR, Colm Rooney GMP, O'Hare, GMP, O'Donoghue, R (1999) What is a social robot?
10. Debarba HG, Bovet S, Salomon R, Blanke O, Herbelin B, Boulic R (2017) Characterizing first and third person viewpoints and their alternation for embodied interaction in virtual reality. PLOS ONE 12(12):1–19. https://doi.org/10.1371/journal.pone.0190109
11. Goldberg K (2001) The robot in the garden: telerobotics and telepistemology in the age of the internet. MIT Press
12. Gomez R, Szapiro D, Merino L, Brock H, Nakamura K, Sabanovic S (2020) Emoji to robomoji: exploring affective telepresence through haru. In: International Conference on Social Robotics. Springer, pp 652–663. https://doi.org/10.1007/978-3-030-62056-1_54
13. Gorisse G, Christmann O, Amato EA, Richir S (2017) First- and third-person perspectives in immersive virtual environments: presence and performance analysis of embodied users. Frontiers in Robotics and AI. 4:33. https://doi.org/10.3389/frobt.2017.00033
14. Higgins I, Matthey L, Pal A, Burgess C, Glorot X, Botvinick M, Mohamed S, Lerchner A (2017) Beta-vae: learning basic visual concepts with a constrained variational framework. Iclr 2(5):6
15. Hoffman G, Ju W (2014) Designing robots with movement in mind. Journal of Human-Robot Interaction 3(1):89–122
16. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Bach F, Blei D (eds) Proceedings of the 32nd international conference on machine learning (proceedings of machine learning research), vol 37. PMLR, Lille, France, 448–456. http://proceedings.mlr.press/v37/ioffe15.html
17. Jonggil A, Kim GJ (2018) SPRinT: A mixed approach to a hand-held robot interface for telepresence. International Journal of Social Robotics 10(4):537–552. https://doi.org/10.1007/s12369-017-0463-2
18. Kilteni K, Groten R, Slater M (2012) The sense of embodiment in virtual reality. Presence: Teleoperators and Virtual Environments 21(4):373–387. https://doi.org/10.1162/PRES_a_00124
19. Kingma DP, Welling M (2013) Auto-encoding variational bayes. arXiv:1312.6114 [stat.ML]
20. Komiyama R, Miyaki T, Rekimoto J (2017) JackIn space: designing a seamless transition between first and third person view for effective telepresence collaborations. In: Proceedings of the 8th augmented human international conference (AH '17). association for computing machinery, New York, NY, USA, Article 14. https://doi.org/10.1145/3041164.3041183
21. Kucherenko T, Jonell P, Yoon Y, Wolfert P, Henter GE (2021) A large, crowdsourced evaluation of gesture generation systems on common data: the GENEA challenge 2020. In: International conference on intelligent user interfaces. https://doi.org/10.1145/3397481.3450692
22. Mandlekar A, Zhu Y, Garg A, Booher J, Spero M, Tung A, Gao J, Emmons J, Gupta A, Orbay E, Savarese S, Fei-Fei L (2018) RoboTurk: A crowdsourcing platform for robotic skill learning through imitation. In: proceedings of the 2nd conference on robot learning (proceedings of machine learning research), vol 87. PMLR, pp 879-893. http://proceedings.mlr.press/v87/mandlekar18a.html
23. Marmpena M, Lim A, Dahl TS, Hemion N(2019) Generating robotic emotional body language with variational autoencoders. In: 2019 8th international conference on affective computing and intelligent interaction (ACII). IEEE, pp 545–551
24. Müller J, Langlotz T, Regenbrecht H (2016) PanoVC: Pervasive telepresence using mobile phones. In: 2016 IEEE international conference on pervasive computing and communications. pp 1–10. https://doi.org/10.1109/PERCOM.2016.7456508
25. Pavllo D, Feichtenhofer C, Grangier D, Auli M (2019) 3D Human pose estimation in video with temporal convolutions and semi-supervised training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)
26. Raaen K, Kjellmo I (2015) Measuring latency in virtual reality systems. Entertainment computing - ICEC 2015. Springer International Publishing, Cham, pp 457–462

27. Rakita D, Mutlu B, Gleicher M (2020) Effects of onset latency and robot speed delays on mimicry-control teleoperation. In: Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction (HRI '20). Association for computing machinery, New York, NY, USA, 519–527. https://doi.org/10.1145/3319502.3374838
28. Rhodin H, Tompkin J, Kim KI, Varanasi K, Seidel H-P (2014) Theobalt C (2014) Interactive motion mapping for real-time character control. Computer Graphics Forum 33(2):273–282. https://doi.org/10.1111/cgf.12325
29. Russell JA (1980) A circumplex model of affect. Journal of Personality and Social Psychology 39(6):1161
30. Sakashita M, Minagawa T, Koike A, Suzuki I, Kawahara K, Ochiai Y (2017) You as a puppet: evaluation of telepresence user interface for puppetry. In: Proceedings of the 30th annual ACM symposium on user interface software and technology (UIST '17). Association for Computing Machinery, New York, NY, USA, 217–228. https://doi.org/10.1145/3126594.3126608
31. Schmidhuber J (2015) Deep learning in neural networks: an overview. Neural Netw 61(2015):85–117. https://doi.org/10.1016/j.neunet.2014.09.003
32. Seol Y, O'Sullivan C, Lee J (2013) Creature features: online motion puppetry for non-human characters. In: Proceedings of the 12th ACM SIGGRAPH/eurographics symposium on computer animation (SCA '13). ACM, New York, NY, USA, 213–221. https://doi.org/10.1145/2485895.2485903
33. Sirkin D, Ju W (2012) Consistency in physical and on-screen action improves perceptions of telepresence robots. In: Proceedings of the seventh annual ACM/IEEE international conference on human-robot interaction (HRI '12). Association for Computing Machinery, New York, NY, USA, 57–64. https://doi.org/10.1145/2157689.2157699
34. Slyper R, Hoffman G, Shamir A (2015) Mirror puppeteering: animating toy robots in front of a webcam. In: Proceedings of the ninth international conference on tangible, embedded, and embodied interaction. ACM, pp. 241–248
35. Stiehl WD, Lee JK, Breazeal C, Nalin M, Morandi A, Sanna A (2009) The huggable: a platform for research in robotic companions for pediatric care. In: Proceedings of the 8th international conference on interaction design and children (IDC '09). Association for Computing Machinery, New York, NY, USA, 317–320. https://doi.org/10.1145/1551788.1551872
36. Strong R, Gaver B (1996) Feather, scent and shaker: supporting simple intimacy. Proceedings of CSCW 96:29–30
37. Suguitan M, Gomez R, Hoffman G (2020) MoveAE: modifying affective robot movements using classifying variational autoencoders. In: Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction. pp 481–489. https://doi.org/10.1145/3319502.3374807
38. Suguitan M, Hoffman G (2019) Blossom: a handcrafted open-source robot. ACM Transactions on Human-Robot Interaction 8(1) Article 2 (March 2019), 27. https://doi.org/10.1145/3310356
39. Taheri S, Beni LA, Veidenbaum AV, Nicolau A, Cammarota R, Qiu J, Lu Q, Haghighat MR (2015) WebRTCbench: a benchmark for performance assessment of webRTC implementations. In: 2015 13th IEEE Symposium on embedded systems for real-time multimedia (ESTIMedia). pp 1–7. https://doi.org/10.1109/ESTIMedia.2015.7351769
40. Tanaka K, Nakanishi H, Ishiguro H (2014) Comparing video, avatar, and robot mediated communication: pros and cons of embodiment. In: Yuizono T, Zurita G, Baloian N, Inoue T, Ogata H (eds) Collaboration technologies and social computing. Springer, Berlin, pp 96–110
41. Tang A, Fakourfar O, Neustaedter C, Bateman S (2017) Collaboration in 360° videochat: challenges and opportunities. https://doi.org/10.11575/PRISM/10182
42. Tsoi N, Connolly J, Adéníran E, Hansen A, Pineda KT, Adamson T, Thompson S, Ramnauth R, Vázquez M, Scassellati B (2021) Challenges deploying robots during a pandemic: an effort to fight social isolation among children. In: Proceedings of the 2021 ACM/IEEE international conference on human-robot interaction (HRI '21). Association for Computing Machinery, New York, NY, USA, 234-242. https://doi.org/10.1145/3434073.3444665
43. Yamane K, Ariki Y, Hodgins J (2010) Animating Non-Humanoid Characters With Human Motion Data. In: Proceedings of the 2010 ACM SIGGRAPH/eurographics symposium on computer animation (SCA '10). Eurographics Association, Goslar Germany, Germany, 169–178. http://dl.acm.org/citation.cfm?id=1921427.1921453
44. Yoon Y, Cha B, Lee J-H, Jang M, Lee J, Kim J, Lee G (2020) Speech gesture generation from the trimodal context of text, audio, and speaker identity. ACM Transactions on Graphics 39(6) Article 222. https://doi.org/10.1145/3414685.3417838
45. Young J, Langlotz T, Cook M, Mills S, Regenbrecht H (2019) Immersive telepresence and remote collaboration using mobile and wearable devices. IEEE Transactions on Visualization and Computer Graphics 25(5):1908–1918. https://doi.org/10.1109/TVCG.2019.2898737
46. Zhang H, Cisse M, Dauphin YN, Lopez-Paz D (2017) Mixup: beyond empirical risk minimization. arXiv:1710.09412 [cs.LG]