Train With Me: A Study Comparing a Socially Assistive Robot and a Virtual Agent for a Rehabilitation Task.

Valentina Vasco*[1], Cesco Willemse*[2], Pauline Chevalier[2], Davide De Tommaso[2], Valerio Gower[3], Furio Gramatica[3], Vadim Tikhanoff[1], Ugo Pattacini[1], Giorgio Metta[1], Agnieszka Wykowska[2]

[1] iCub, Istituto Italiano di Tecnologia (IIT), Via San Quirico 19D, 16163, Genoa, Italy
{name.surname}@iit.it
[2] Social Cognition in Human-Robot Interaction, Istituto Italiano di Tecnologia (IIT), Via Enrico Melen 83, 16153, Genoa, Italy
{name.surname}@iit.it
[3] IRCCS Fondazione Don Carlo Gnocchi, Via Capecelatro 66, 20148, Milan, Italy
{vgower,fgramatica}@dongnocchi.it

* Both authors contributed equally to this manuscript.

Abstract. Long-term motor deficits affect approximately two thirds of stroke survivors, reducing their quality of life. An effective rehabilitation therapy requires intense and repetitive training, which is resource demanding. Virtual Agents (VAs) and Socially Assistive Robots (SARs) offer high intensity, repetitive and reproducible therapy and are thus both promising as rehabilitation tools. In this paper, we compare a SAR and a VA during a rehabilitation task in terms of users' engagement and movement performance, while leveraging neuroscientific methods to investigate potential differences at the neural level. Results show that our participants' performance on the exercise was higher with a SAR than with a VA, which was especially clear under conditions of decreased perceptual information. Our participants reported higher levels of engagement with the SAR. Taken together, we provide evidence that SARs are a favorable alternative to VAs as rehabilitation tools.

Keywords: Socially assistive robot, Virtual agent, Embodiment.

1 Introduction

According to the World Health Organization (WHO), 15 million people suffer stroke worldwide yearly and, among the survivors, approximately 75% exhibit persistent upper extremity deficits, as limb weakness and impairment of grasping movements [1]. A substantial number of activities of daily living involve the use of the upper limbs and thus such disabilities can severely affect the quality of life.

Intensive and repetitive therapeutic training can significantly reduce motor impairment and lead to a partial or complete motion recovery, as patients re-learn the kinetic movements of the affected limbs [2]. However, such rehabilitation process requires supervision of trained professionals, with a consequent increase of the workload of therapists, whose number is not sufficient to accommodate such needs, both in and (especially) out of the clinic. Innovative solutions could be adopted to efficiently and effectively respond to this demand, with the aim of augmenting current care standards while allowing for greater flexibility of both patients and therapists. An area of active research is that of Virtual Agents (VA), which simulate a physiotherapist in a virtual environment: therapy costs can be reduced as patients can perform the exercises off site (e.g. in their home) and the "gamified" training activity improves their motivation and engagement, with the potential of high deployability and low maintenance [4]. Socially Assistive Robots (SAR) have also recently emerged, with the focus of aiding humans through social interaction, rather than offering only physical support [3]: a SAR guides the user in accomplishing a task through non-contact feedback, encouragement and constant monitoring. Both solutions have the potential of delivering high intensity and repetitive therapy, while providing a reliable and reproducible way to measure improvements in performance. With respect to a VA, a SAR offers physical embodiment and presence, which can positively affect the patient's motivation in terms of persuasion and attraction [5], but also performance (e.g. persistence in performing the exercise [4]).

Previous research has demonstrated the efficacy of social robots in a number of domains, including elderly care [5], daily activities [17], physical therapy [4,6] and stroke rehabilitation [7]. On the other hand, virtual agents have been shown to provide a viable alternative in the same domains [8,9]. Therefore the question of whether to implement a robotic or a virtual agent remains open. Several works directly compare the effect of a SAR and a VA on the user. Schneider et al. [4] analyze data from previous studies to investigate the effect of an embodied robot (the SoftBank Robotics Nao) on users' motivation, compared to a video of a human performing the exercise. Results show that participants training with a robot exercise significantly longer than with a virtual partner and that the robot elicits at least the same motivational effect as the VA. Fasola et al. [5] evaluate the role of physical embodiment by comparing the effect of a robot (the BlueSky Robotics Bandit) to its virtual counterpart. The study shows a strong user preference for physical robot embodiment over the virtual counterpart. Results are further confirmed in [6], which shows a dependence on the use of an embodied agent, as opposed to a simulated agent, when assessing adherence to a physical therapy. A complete review can be found in [10].

In general, previous works show that the physical embodiment and presence of a robot are beneficial to user interaction, in terms of persuasiveness and task performance. In this paper, we compare a SAR and its virtual version displayed on a screen during a typical motor rehabilitation task, consisting of a left shoulder abduction, while the real and the virtual agent monitor, assist and encourage the participants. The aim of the study is to investigate whether participants respond differently to a session with the robot and its virtual version in terms of task performance and engagement. We also adopt neuroscientific methods to investigate the potential differences that occur at neural level.

2 The framework

In this paper, we used the humanoid robot R1 [12] and developed its virtual version within the simulation environment Gazebo. The devised framework, shown in Fig. 1, consists of a set of modules interconnected on a YARP network and is detailed in the next Sections.

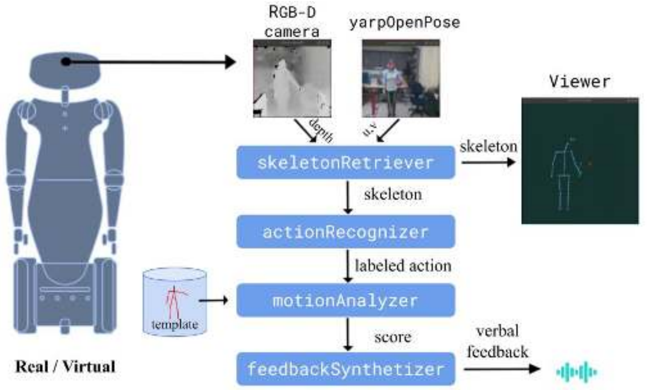

Fig. 1. The devised framework.

2.1 3D skeleton acquisition

yarpOpenPose estimates human poses based on *OpenPose* [11], an open-source library for real-time multi-person 2D pose estimation. The pipeline processes an RGB image and outputs a list of 2D keypoints for each person found in the image, achieving high accuracy and speed regardless of the number of people present in the image. *skeletonRetriever* combines the 2D locations of the keypoints with the depth provided by the camera to reconstruct keypoints in 3D fashion, using the classical pinhole camera model.

2.2 Motion evaluation and feedback

The motion evaluation technique is based on two interconnected layers: an action recognition layer (i.e. *actionRecognizer*), which classifies the skeleton data in a predefined temporal interval, and a subsequent motion analysis layer (i.e. *motionAnalyzer*), which further analyzes the skeleton joints if the action predicted has the same label as that of the exercise to perform (i.e. *abduction*). A feedback synthesizer layer (i.e. *feedbackSynthetizer*) finally transforms the numeric analysis into a verbal feedback with instructions to improve the execution of the exercise.

Action recognition. This layer aims at preventing an erroneous analysis when the motion repertoire includes a high number of exercises, which can be very similar to each other (for example abduction and rotation of the shoulder have common moving joints). In such case, modeling the exercise as a set of interconnected static/dynamic joints is not feasible as covering all the possible joint configurations would be very complex and prone to errors. Furthermore, random movements can not be modeled *a-priori* and would thus lead to wrong analysis.

Action recognition is carried out using a Recurrent Neural Network with Long Short-Term Memory (LSTM) cells [13]. We trained the network offline in a supervised fashion, providing paired input (2D skeleton joints of the upper body provided by *skeletonRetriever*) and output (label of the exercise). Specifically, the input consists of temporal sequences of fixed length of the skeleton joints, which, at training time, include different parts of the movement, such that the classification is not dependent on a specific section of the movement. The training set was collected recording from a frontal and a side camera view 5 repetitions of the exercise, with the single movement performed 10 times. For the experiments presented in the paper, the network was designed to have two classes, namely a shoulder abduction and random movements, but can be easily extended to include more exercises.

Motion analysis. This layer compares the observed skeleton with a pre-recorded template moving coherently with the robot. Spatial alignment between the two is achieved using the roto-translational offset between the bodies, extracted by the shoulders and the hips 3D positions. Temporal alignment is also achieved applying Dynamic Time Warping (DTW) to the 3D joint positions to extract the optimal warping path *w* between candidate and template joints. We then compute the error in position ε between each component of candidate *c* and template skeletons *t* in a predefined temporal window, as following:

$$\varepsilon_i = \sum_{k=0}^{N} (c_i(w_k) - t_i(w_k)),$$

with $i = \{x, y, z\}$ and *N* indicating the joint component and the temporal window's length. Positive (negative) tails in the error distribution reflect a *c* with a higher (lower) range of motion than *t* and are identified applying a threshold $\overline{\gamma}$ to the skewness γ:

$$\gamma_i = E\left[\left(\frac{\varepsilon_i - \mu_i}{\sigma_i}\right)^3\right] > \overline{\gamma}\ (< -\overline{\gamma}),$$

with (μ_i, σ_i) being the distribution mean and standard deviation.

Finally, we evaluate the speed performing the Fourier transform of each component of the joints under analysis in the defined temporal window. A difference in frequency can be related to a difference in speed, reflecting a skeleton moving slower (faster) than the template if positive (negative).

Based on the detected errors, each participant is associated to a score in a range of [0, 1], with 0 indicating a completely wrong movement and 1 a movement performed perfectly. Values in the middle reflect an error in speed or in position.

Verbal feedback. This layer is responsible for providing a real-time verbal feedback to the participants, according to the strategy summarized in Table 1.

Table 1. Verbal feedback.

Detected error	Score	Verbal feedback
Action not recognized	0.0	Please put more effort.
df > 0 (< 0)	0.5	Move the left arm faster (slower).
$\varepsilon_x > 0$ (< 0)	0.5	Move the left arm more on the left (on the right).
$\varepsilon_y > 0$ (< 0)	0.5	Move the left arm further up (further down).
$\varepsilon_z > 0$ (< 0)	0.5	Move the left arm backward (forward).
–	1.0	You are moving very well.

3 Experimental design

We used convergent methods to systematically validate the framework and to compare user engagement with an embodied (SAR) versus the virtual version (VA) of R1 whilst doing the exercise.We investigated self-reported engagement as well as performance measures under different conditions: observation, visible imitation, and occluded imitation. During the observation phase, participants merely observed the agent demonstrating the abduction movement. This condition was primarily used to establish baseline electroencephalography (EEG). During the visible imitation condition, participants executed the movement together with the robot, much like in a realistic rehabilitation exercise scenario, during which verbal feedback was provided by the agent at two moments. Finally, during the occluded imitation condition, the arm of the robot was removed from the view whilst the participant attempted to continue the movement in synchrony with the agent. To enhance the effects of absence of information, verbal feedback was not provided in this condition. We employed this condition for two reasons. First, this allowed us to measure real-life performance in situations where perceptual information processing is further from perfect, but more importantly, this allowed us to examine how well the movement was maintained in mental representation as a measure of engagement.

3.1 Materials

To compare the engagement between the SAR and VA, we measured self-reported engagement (questionnaires), a collection of performance metrics, and EEG.

Self-report questionnaires. We used a 10-item questionnaire to evaluate the participant's engagement with the experiment, adapted from the previous literature [14,15]. The items were scored on 7-point Likert-scale (1 – "not at all", 7 – "very much"), see Table 2 for the items.

Table 2. Self-report questionnaire item completed after the SAR and VA sessions.

Nr	Item question	Nr	Item question
1	How engaging was the interaction?	6	I was so involved in the interaction that I lost track of time.

2	How relaxing was the experience?	7	Overall, to what extent do you think you were able to engage with R1?
3	How exciting was the experience?	8	Overall, to what extent would you say that you liked R1?
4	How completely were your senses engaged?	9	To what extent do you feel you have developed a relationship with R1?
5	The experience caused real feelings and emotions for me.	10	How engaged do you think R1 was with you?

Performance metrics. As a more objective indication of engagement, we compared the following metrics of the participant's arm movements during the imitation and occlusion conditions against the template. *Hand position:* The cross-correlation between the participant's hand position (X & Y) and the template for each condition to determine the delay between the two signals. We corrected the lag between the two signals and computed their correlation coefficients *Amplitude:* The euclidean distance between the participant's and the template's mean peak to peak vertical distance. *Feedback:* Mean participant scores as outlined in Table 1.

EEG. We recorded EEG to examine the potential of recording human neuronal activity in naturalistic human-robot interaction – as opposed to typical lab-based psychology experiments during which participants are explicitly asked to sit still and observe well-controlled stimuli on a screen. More specifically, we examined the mu-rhythm, which is typically observed on electrodes placed over the sensorimotor cortex under rest. Mu-rhythm (typically 8-13 Hz) has been initially observed as being suppressed under execution of motion, but later studies also showed mu-suppression for mere observation [16]. We examined how mu suppression changes between conditions of observation of movement and executing movement. Specifically, this served as a manipulation check to see if performance differences could not be attributed to differential motion processing. To these means, we used a 16-channel setup (BrainProducts ActiCap and V-Amp) in which 15 active electrodes were positioned on the scalp covering the midline. We registered horizontal and vertical eye movements with three dedicated electrodes.

3.2 Participants and procedure

Sixteen participants took part (age M=23.0, S.D.=2.81, 7 males). After giving informed consent, we fitted them with the EEG equipment, gave task instructions, and gave them ear plugs to wear (to attenuate background and actuator noise). This study was approved by the local Ethics Committee (Comitato Etico Regione Liguria).

The experimental procedure was as follows (cf. Figure 3). Whilst sat on a chair, participants first only observed R1 executing the abduction movement (observation), after which they would be asked to do the movement together (visible imitation). Then, R1's arm would be occluded by an experimenter placing a panel in front of the robot's shoulder joint with the SAR, or by presenting a virtual panel on the screen (occluded imitation). Each of these conditions lasted for eight arm movements and the total sequence was repeated six times to increase statistical power. Next, participants completed a questionnaire about their engagement during the previous session and repeated this entire procedure with the VA if they started with the SAR, or with the SAR if they started with the VA (test order was counterbalanced between participants).

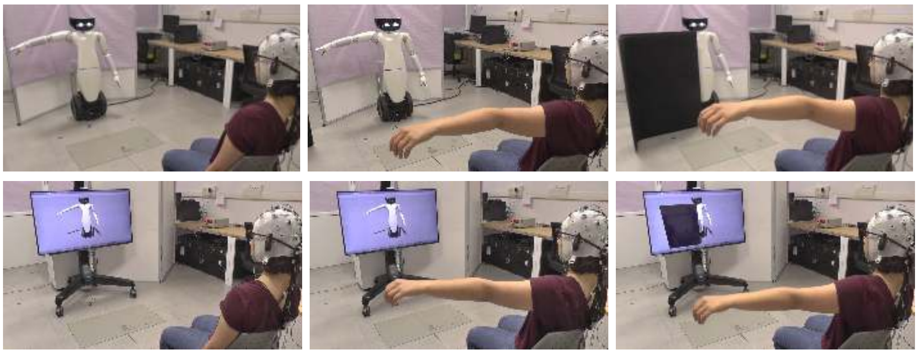

Fig. 3. The experimental scenarios during observation, visible imitation and occluded imitation (first, second and third column) for embodied (first row) and virtual agent (second row).

4 Results

4.1 Self-report questionnaires

We found a significant difference in self-reported user engagement between the SAR and VA, $t(15)=4.37$, $p<.001$. These data showed that the interaction with the SAR was more positively rated (M=43.8, S.D.=12.2) than with the VA (M=33.9, SD=13.1). This pattern was also observed when we analyzed items individually with Wilcoxon signed-rank tests, except for items 2, 5, and 6; *p*s≥.39.

4.2 Performance metrics

The performance metrics below were subjected to 2 (agent; SAR, VA) X 2 (condition; visible imitation, occluded imitation) repeated-measures ANOVAs.

Hand position. For both the X and Y-positions, we found main effects of agent (X: $F(1,15)=34.9$, $p<.001$, Y: $F(1,15)=41.8$, $p<.001$. Participants had better performance with the SAR) and of condition (X: $F(1,15)=43.7$, $p < .001$, Y: $F(1,15)=28.0$, $p<.001$. Performance was better during the visible imitation condition than during occlusion). More critically, we also found and interaction effect of agent and condition (X: $F(1,15)=31.9$, $p<.001$. Y: $F(1,15)=28.0$, $p<.001$). For the visible imitation condition, there was no significant difference between the embodied and the virtual agent. However, during the occluded imitation condition, participants performed better with the embodied agent than with the virtual one (X and Y: *p*s<.001). Further, with the virtual agent, the imitation phase was better performed than the occlusion phase (X and Y: *p*s<.001), whereas no significant difference was found between the imitation and occlusion phase for the embodied agent (see Figure 4).

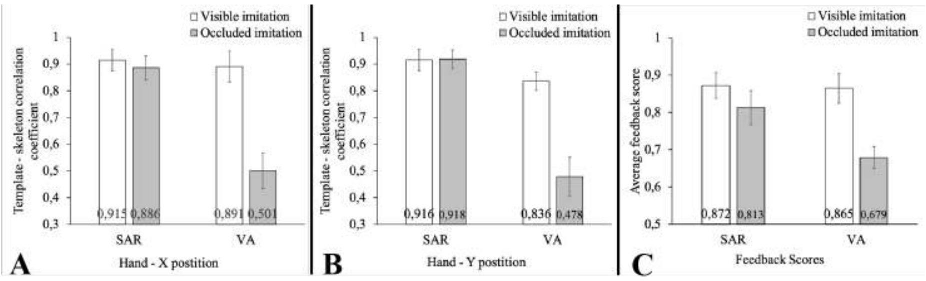

Fig. 4. A: Correlation coefficients between the participants and the template for the left hand's X-position and **B:** Y-position. **C:** Average feedback scores per condition. Error bars: +/- 1 SE.

Amplitude. Here, we found a main effect of condition in X-position (*F*(1,15)=5.80, *p*=0.029). Compared with the visible imitation condition (M=0.038, S.D.=0.019), the deviation of the amplitudes from the template was higher in occluded imitation condition (M=0.042, S.D.=0.020) in the X-position. No significant effects were observed in Y-position.

Feedback scores. We found a main effect of condition, *F*(1,15)=29.3, *p*<.001. Participants performed worse when the robot's arm was occluded compared to visible imitation. There was no main effect of agent (*p*=.14), but an agent x condition interaction effect emerged; *F*(1,15)=7.70, *p*=.014. Paired samples t-tests revealed that whereas performance was worse during occlusion for both agents (SAR *p*=.016, VA *p*<.001), participants performed 13.4% better (cf. Figure 4) when the SAR was occluded compared with the VA occlusion, *t*(15)=2.77, *p*=.014. There was no difference in performance between the agents during the visible imitation condition (*p*=.90). Taken together, the performance metrics demonstrate that participants performed the exercise better, and were thus more engaged, with the SAR than with the VA, especially when hindered perception is taken into account.

4.3 **EEG**

The EEG data were filtered (IIR Butterworth, 0.5 Hz – 80 Hz 24 dB/oct, 50 Hz notch) before carrying out a Gratton and Coles ocular correction. These data were then segmented, and other artifacts were rejected semi-automatically. We then carried out a Fast Fourier Transformation. For analysis we exported mean activity (µV) between 8.0–13.0 Hz and calculated the average value of the electrodes where mu-rhythm is typically observed, C3/C4. On these data, we conducted a 2 (agent: SAR, VA) X 3 (condition: observation, imitation, occlusion) repeated measures ANOVA.We excluded one participant from the analyses because the mu-rhythm was not observed in the individual data.

We found a main effect of condition, $F(2,26)=11.4$, $p <.001$. Post hoc comparisons revealed that with both agents more mu-rhythm was evoked during the observation phase than during the imitation phase ($p<.001$) and than during the occlusion phase ($p=.003$). Mu-rhythm during visible imitation did not differ significantly between visible imitation and occluded imitation (p=.37). We found no main effect of agent nor an interaction effect. See Figure 5 for scalp distributions of mu-rhythm.

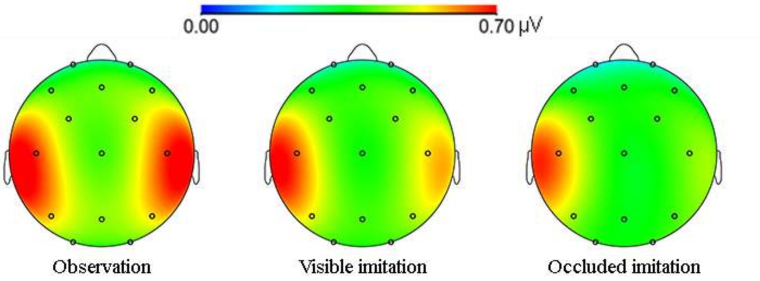

Fig. 5. Power (µV) and scalp distribution of mu-rhythm (8.0 - 13.0 Hz) in the three different conditions, collapsed across both agents.

5 Discussion

In summary, our questionnaire and performance data both show that engagement was higher with the SAR compared with the VA. The EEG data confirm that mu-suppression with an artificial agent shows a similar pattern as often observed for human movements, whilst demonstrating the feasibility of future mu-rhythm studies in more naturalistic setting than well-controlled experiments with static stimuli. Moreover, mu-rhythm was indeed equally suppressed for both agents between the visible imitation and occluded imitation conditions, so the performance results cannot be attributed to differences in motor activity *per se*.

Our results support the idea that a SAR as rehabilitation tool improves both participants' engagement and performance, if compared to a VA. This is in line with previous research [4–6,10], but uniquely, the current study expands on the literature by combining questionnaire data with 1) direct performance metrics, 2) assessing the mental representation of the interacting agent by subtracting perceptual information from the scenario (occlusion condition) and 3) using EEG to check that the findings can not be attributed to different motion processing. Furthermore, we used structured repetition of trials to increase reliability of our measures.

Importantly, although the rehabilitation task proposed was quite easy, the level of engagement with the SAR was already significantly higher than with a VA. We plan to design a more complex setup with the participant having to touch the agent's hand, which also navigates towards him/her. Such task requires a deeper level of interaction and thus might further enhance the user's engagement with the real robot.

In conclusion, we propose that a robot's physical presence engages humans in exercise more compared to screen-based animations. Robot's presence increases motivation leading to better performance, potentially aiding in a faster recovery.

Acknowledgments. This project is funded by the Joint Lab between IIT and Fondazione Don Carlo Gnocchi Onlus and by the Minded Program - Marie Skłodowska-Curie grant agreement No 754490 (fellowship awarded to Pauline Chevalier). The authors thank Nina Hinz and Serena Marchesi for their assistance with data collection.

References

1. Lawrence E.S. *et al.*: Estimates of the Prevalence of Acute Stroke Impairments and Disability in a Multiethnic Population. Stroke (5), 1279–84 (2001).
2. Carlsson H. *et al.*: SENSory re-learning of the UPPer limb after stroke (SENSUPP): Study protocol for a pilot randomized controlled trial. Trials (19) (2018).
3. Feil-Seifer D., Mataric M.J.: Defining socially assistive robotics. In: Proceedings of the 2005 IEEE 9th Int. Conf. on Rehabilitation Robotics, pp. 465–468 (2005).
4. Schneider S., Kummert F.: Comparing the effects of social robots and virtual agents on exercising motivation. In: 10th Int. Conf. on Social Robotics ICSR, pp. 451–461 (2018).
5. Fasola J., Mataric M.: Comparing physical and virtual embodiment in a socially assistive robot exercise coach for the elderly. Tech. Rep. CRES–11–003 (2011).
6. Brooks D., Chen Y., Howard A.: Simulation versus embodied agents: Does either induce better human adherence to physical therapy exercise? In: Proceedings of the IEEE Int. Conf. on Biomedical Robotics and Biomechatronics, pp. 1715–1720 (2012).
7. Mataric M. *et al.*: Socially assistive robotics for stroke and mild TBI rehabilitation. Studies in health technology and informatics (145), 249–62 (2009).
8. Fiol-Roig G. *et al.*: The Intelligent Butler: A Virtual Agent for Disabled and Elderly People Assistance. In: Int. Symp. on Distributed Computing and Artificial Intelligence 2008. Advances in Soft Computing, vol 50. Springer, Berlin, Heidelberg (2008).
9. Arip E.S.M. *et al.*: Virtual reality rehabilitation for stroke patients: Recent review and research issues. AIP Conference Proceedings. 1905. 050007. 10.1063/1.5012226.
10. Li J.: The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. Int. Journal of Human Computer Studies (77), 23–37 (2015).
11. Cao Z. *et al.*: OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields, arXiv preprint arXiv:1812.08008 (2018).
12. Parmiggiani A. *et al.*: The design and validation of the R1 personal humanoid. In: 2017 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS), pp. 674–680 (2017).
13. Chevalier G.: LSTMs for Human Activity Recognition, 2016 https://github.com/guillaume-chevalier/LSTM-Human-Activity-Recognition.
14. Sidner C.L. *et al.*: Explorations in engagement for humans and robots, Artificial Intelligence, 166(1-2), 140–164 (2005).
15. Hall J. *et al.*: Perception of own and robot engagement in human–robot interactions and their dependence on robotics knowledge, Robotics and Autonomous Systems, 62(3), 392–399 (2014).
16. Hobson H.M., Bishop V.M.: The interpretation of mu suppression as an index of mirror neuron activity: past, present and future, Royal Society Open Science, 4(3) (2017).
17. Rossi S. *et al.*: Socially Assistive Robot for Providing Recommendations: Comparing a Humanoid Robot with a Mobile Application, Int. Journal of Social Robotics (10), 265–278 (2018).