Gesture recognition with a 2D low-resolution embedded camera to minimise intrusion in robot-led training of children with autism spectrum disorder

ERCOLANO, Giovanni, ROSSI, Silvia, CONTI, Daniela <http://orcid.org/0000-0001-5308-7961> and DI NUOVO, Alessandro <http://orcid.org/0000-0003-2677-2650>

Applied Intelligence (2024) 54:6579–6591
https://doi.org/10.1007/s10489-024-05477-z

Gesture recognition with a 2D low-resolution embedded camera to minimise intrusion in robot-led training of children with autism spectrum disorder

Giovanni Ercolano[1] · Silvia Rossi[1] · Daniela Conti[2] · Alessandro Di Nuovo[3]

Accepted: 18 April 2024 / Published online: 20 May 2024

Abstract
Growing evidence shows the potential benefits of robot-assisted therapy for children with Autism Spectrum Disorder (ASD). However, when developing new robotics technologies, it must be considered that this condition often causes increased anxiety in unfamiliar settings. Indeed, children with ASD have difficulties accepting changes like introducing multiple new technological devices in their routines, therefore, embedded solutions should be preferred. Also, in this context, robots should be small as children find the bigger ones scary. This leads to limited computing resources onboard as small batteries power them. This article presents a study on gesture recognition using video recorded only by the camera embedded in a NAO robot, while it was leading a clinical procedure. The video is 2D and low quality because of the limits of the NAO-embedded computing resources. The recognition is made more challenging by robot movements, which alter the vision by moving the camera and sometimes by obstructing it with the robot's arms for short periods. Despite these challenging real-world conditions, in our experiments, we have tuned and improved state-of-the-art algorithms to yield an accuracy higher than 90% in the gesture classification, with the best accuracy being 94%. This level of accuracy is suitable for evaluating the children's performance and providing information for the diagnosis and continuous assessment of the therapy. We have also considered the performance improvement of using a low-power GPU-AI accelerator embedded system, which could be included in future robots, to enable gesture analysis during the therapy, which could be adapted to the child's performance.

Keywords Autism spectrum disorder · Intellectual disability · Gesture recognition · Deep learning · Automatic feature extraction

1 Introduction

Interdisciplinary research is successfully exploring robotic technologies as personalised social companions to deliver or supplement behavioural interventions [1, 2]. Socially Assistive Robots (SARs) stand out as the most sophisticated among emerging robotics technologies. These robots incorporate audio, visual, and movement interfaces, as well as embedded computing hardware for edge AI, to simulate social behaviour such as complex dialogue with non-verbal communication, recognising emotions, and physical interaction with humans [3]. The primary objective of SAR is to establish a positive and productive interaction with humans, while also providing assistance and improving their quality of life. These robots are often utilized in domains such as motivation, rehabilitation, or learning, with the goal of achieving measurable progress in these areas [4]. SAR provides a physical manifestation for intelligent agents, rather than being confined to a digital screen. This means that SARs can be present in the physical world and directly interact with humans and objects in their environment [5]. They are capable of engaging with users through a rich variety of sensory modalities such as sound, sight, and touch. This allows for multiple options for delivering content or interactions, which can be customized to improve their effectiveness based on individual user preferences or physical abilities [6]. Several studies showed that SARs can support therapy and training of children with Autism Spectrum Disorder (ASD), who have difficulties in social interaction because of their condition,

✉ Alessandro Di Nuovo
a.dinuovo@shu.ac.uk

1 University of Naples Federico II, Napoli, Italy

2 University of Catania, Catania, Italy

3 Sheffield Hallam University, Sheffield, UK

which has a male-to-female prevalence of 4:1 [7] . Children with ASD consider robots' behaviour more predictable than humans and, therefore, it is easier for them to accept robots as social partners [8]. SARs can prompt children with ASD in a realistic social interaction via their physical presence and simulated social abilities, including non-verbal cues like eye gaze, gestures, and posture [9]. Indeed, many clinical studies [10] demonstrated significant benefits in the treatment of children with ASD, e.g. they can enhance training [11] and perform automated assessment [12]. ASD is a difficult condition, which also includes difficulties in processing novelty, which can cause anxiety and negative responses by an individual with ASD [13]. In this context, it is of fundamental importance to limit the introduction of novel technological devices to the strictly necessary ones. Indeed, children with ASD can be upset by the introduction of many novel items in their environment, therefore, simplicity is an essential prerequisite for successfully including new technology in the therapy for the widest range of children with ASD. To monitor and acquire information from the interaction, the use of bulky external setups (e.g. computers, multiple cameras and other devices) should be avoided, as they can cause distress to the child. The best approach is to use only the robot's embedded sensors and computing abilities to record data [14]. This necessity represents a challenge for the application of SAR in real contexts like clinical therapy because the onboard computing of commercial robotic platforms is limited to account for multiple constraints such as cost, space, heat, and power consumption. In fact, commercial platforms that are commonly used with ASD, do not have sufficient computing resources to concurrently control the robot and acquire data from sensors during the clinical interaction.

In our clinical studies, we minimise the intrusion in the therapeutic setting to avoid upsetting the children. In this article, we investigate the feasibility and propose a proof-of-concept prototype of automatic gesture recognition using only the data collected by the robot's embedded camera without the use of any other device. The clinical study in which we collected the data consisted of robot-assisted imitation training (see Fig. 4) with six male children (M-chronological age=104.3 months, range=66-121, SD=18.6) with ASD and ID. Two children had a profound ID level, two severe ID levels, one moderate ID level and one with a mild ID level. The robot used in the study was the Aldebaran Robotics NAO [15], which is the most common humanoid platform employed in SAR [16]. NAO was used in 80% of studies in which a humanoid was employed for robot-led therapy of children with ASD [17]. The clinical activities included six encounters, in which the NAO robot was prompting the children in three Gross Motor Imitation (GMI) tasks. For each child, the robot's camera recorded the video of 18 procedures (6x3) in total. The robot initiated the procedure by verbally instructing the child with simple and concise language, followed by prompting the child to imitate its movements. Each session lasted around 6-8 minutes per child, with a 1-minute break between each activity to allow the children to rest in the nearby multi-sensory area. More detailed information on the clinical experiment can be found in [18], which provides the details on the methodology and the evaluation of a robot-assisted imitation therapy for children with ASD and Intellectual Disability (ID). During the therapy sessions, the children's imitation of the robot's gestures was recorded to evaluate their performance and track their progress over time. The recordings were manually analysed and labelled accordingly to identify the gestures that children were performing in each frame. These labelled frames form the dataset used in this article.

However, while on one hand, the use of the embedded camera facilitates the acceptance of the system by the children, on the other, this creates a technological challenge because, as common for many commercial robots, the embedded camera does not have the depth measurement and it was only able to record images at a frequency of 10 fps and a resolution of 320x240 pixels because of the limitations of the onboard computational resources (CPU and memory), which were also used to control the robot behaviour for the therapy. This is a common issue with the small robotic platforms that are being used for robot-assisted therapy, which have usually limited computing and sensing on-board. Indeed, the actual resolution and frame rate of cameras could be higher but it is are usually restricted due to the limited computing capacity of the main processor and memory resources [12]. When working with children, particularly those with ID, it can be difficult to enforce constraints that are necessary for optimizing algorithm performance. As a result, it is crucial to be able to accurately estimate a child's visual movements without relying on constraints like confining them in specific positions. While such devices can improve performance, they can also limit the portability of the system and complicate its integration into a standard therapeutic environment.

The unique contribution of this article can be summarised as follows:

- Novel application of machine learning techniques for automated gesture recognition with real-world data, which was collected during robot-led imitation therapy sessions for children with autism spectrum disorder and intellectual disability.
- Identification of optimal parameters for a multi-layer LSTM architecture to maximise accuracy for the assessment of children's success in therapy.
- Proof-of-concept evaluation of a low-power commercial embedded system for edge-AI (NVIDIA Jetson) as a potential solution for real-time computation onboard future robotic platforms.

The rest of the article is organised as follows: Section 2 presents an overview of recent results in gesture recognition applied to human-robot interaction; Section 3 provides the details of the machine learning approaches that were evaluated in our computational experiments with the children dataset described above; Section 5 discusses the results; finally, Section 6 gives our conclusion.

2 Review in gesture recognition during Human-Robot Interaction

There are numerous methods of classification of gestures in the literature. In general, the techniques differ from different feature extraction to classification methods.

Many works involve gesture recognition with OpenPose, manual feature selection and classical machine learning algorithms. In [19], the authors extracted the human pose using OpenPose and recognising the gestures with Dynamic Time Warping (DTM) and One-Nearest-Neighbor (1NN) from the time-series. Other works use instead more devices to better identify gestures. In [20], they obtained 3D skeletal joint coordinates from 2D skeleton extraction with OpenPose and the depth from a Microsoft Kinect 2. Then, the 3D coordinates are used to detect the gesture using a CNN classifier. This system was employed for real-time human-robot interaction.

The gestures can be classified as static or dynamic. A gesture is static if the user assumes a certain pose while it is dynamic when the gesture consists of several poses. For this reason, the identification of gestures is not trivial and also requires temporal segmentation. Classic gesture recognition methods are based on HMM, particle filtering and condensation algorithm, FSM approach, Artificial Neural Networks (ANNs), genetic algorithms (GAs), fuzzy sets and rough sets. Deep neural networks have become state-of-the-art in Computer Vision and are also applied in the recognition of gestures outperforming the previous state-of-the-art methods. For a recent review of classic and deep learning techniques see [21].

In [22], the Authors used OpenPose to capture the 2D positions of a person's joints to compare gesture imitation with recorded gestures. Their goal is to estimate whether real-time movements correspond precisely with standard gestures. To make this comparison they used videos of Tai Chi teaching. From the joints, they calculated the movement trajectory for each point. A similarity metric was defined as the distance between the movement trajectories of the standard and real-time videos. Important features to better describe the gestures are redundancy reduction, robustness, invariance with respect to sensor orientation, signal continuity, and dimensionality reduction. To make the system robust, they defined the trajectory equation with Bézier curves that are robust to input noise. To define the distance between the recorded gesture and the imitated gesture, they calculated the discrete Fréchet distance. From the joints of the trajectories they then obtained 12 distances that composed a vector, finally obtaining a score by applying a weighted distance formula.

In [20] they obtained 3D skeletal joint coordinates from 2D skeleton extraction with OpenPose and the depth from a Microsoft Kinect 2. Then, the 3D coordinates are used to detect the gesture using a CNN classifier. This system was employed for real-time human-robot interaction. Human gesture and activity recognition are some of the main topics of human-machine interaction. Consequently, there are many works in literature. In [23], the authors used the difference between subsequent frames from the depth image of the Microsoft Kinect to recognise eight gestures: CLAP, CALL, GREET, WAVE, NO, YES, CLASP, REST

In [24] a simultaneous gestures system for multiple users was introduced and the results on a maximum of six users had an accuracy higher than 90%. In [25] a Wi-Fi-based zero-effort domain gesture recognition system (Widar3.0) estimates the velocity profiles to characterise the gesture kinetic features. A deep learning model exploits spatial-temporal features for gesture recognition. The accuracy result achieved is high, near 90.0%, independently from the domain in real environments.

In [19] highlighted the need to communicate with service robots through gestures, for example, to draw the robot's attention to someone or something. To avoid using special hardware they used only RGB videos, extracting the pose in the frames of the videos with OpenPose. They present a method for gesture recognition, starting from the pose extracted with OpenPose, in conjunction with Dynamic Time Warping (DTW) and One-Nearest-Neighbor (1NN) for time-series classification. Before passing the joint coordinates to the DTW classifier, the key points are normalised to achieve scale and translation invariance so that they are not dependent on the relative position of the person with respect to the camera. One of the main advantages of this approach is the ability to easily add new gestures. To reduce the number of signals processing by DTW, they considered signal variance. All signals with a low variance, thus indicating no motion, were considered to be uninformative. For classification with 1NN they used warping distance as a metric instead.

In [26] they propose an approach based on the temporal and spatial relationship between joints and joint pairs. To alleviate the variation of the temporal sequence they propose a new temporal transformation module (TTM). Finally, all extracted features are merged into a multi-stream architecture and then classified by a full-connected layer. This kind of

approach has been tested on datasets such as ChaLearn 2013, ChaLearn 2016 and MSRC-12 obtaining very good results.

EfficientGCN-B4 [27], an action transformer, used a fully self-attentional architecture. The skeleton poses are extracted from 2D videos with Openpose [28]. Similar to BERT and Vision Transformers, the sequences are represented as embeddings. The embeddings are fed to a Transformer Encoder. The output is fed into a linear classification head. It exceeds more elaborated networks that mix convolutional, recurrent and attentive layers. The accuracy of the EfficientGCN-B4 [27] outperforms other models like MS-G3D (J+B) [29], MS-G3D (J) [29], ST-TR [30] on MPOSE2021 dataset.

In recent years, there has been a surge of interest in developing accurate and efficient methods for gesture recognition. A number of research papers have been published, each proposing different approaches to address this challenging problem. Some of the most promising methods include Convolutional Transformer Fusion Blocks [31], Spike representation of depth image sequences with spiking neural networks [32], and Deep Hybrid Models that combine Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks [33].

One common theme among these papers is the focus on hand gesture recognition. While accurate hand gesture recognition is undoubtedly important, there are other factors to consider as well. For example, our recent paper on the recognition of gestures in autistic children takes a more holistic approach by considering the entire body of the child as they attempt to imitate the gestures of a humanoid robot.

In this work, we investigate the automatic recognition of gestures using only the RGB camera of the robot's forehead using the video recordings collected during the previous study. We should point out that, even if the video quality of the NAO camera is very low because of the limited computing resources, it has been demonstrated that it is possible to successfully extract the skeleton joints using OpenPose [28], even in case of occlusions [34], with very good accuracy.

3 Methods

In our work, we wanted to automate and make more objective the assessment of the success or failure of the children's imitation of the robots' gestures in a clinical setting. To this end, we investigated the use of neural networks, particularly the Long Short Term Memory (LSTM) recurrent networks, for gesture recognition in a clinical setting. The approach is divided into two steps: first to automatically extract the human skeleton pose and the temporal features between the different poses from a low-res video sequence taken during the clinical therapy; second, the resulting features are classified into the possible gestures; finally, the gesture recognised is compared with the one performed by the robot to assess the success or failure of the imitation.

To make it applicable in real clinical settings, our approach aims at yielding the built-in camera of the NAO robot to recognise the child's gestures from a sequence of 2D poses with a deep recurrent neural network made of 2 LSTMs. This approach reduces to the minimum the intrusion in the child space, which makes it more acceptable and suitable for the real world application than previous experimental settings, e.g. [11].

For the feature extraction we selected OpenPose [28] algorithm (version 1.7.0), the first real-time multi-person system to jointly detect the human body and more, which is the state-of-the-art algorithm to extract the human pose from the image frame of the videos. OpenPose uses a bottom-up approach and it has a constant runtime compared to Alpha-Pose [35–37] (top-down approach) and Mask R-CNN [38] (similar to the top-down approach). It can achieve better accuracy results with different average confidence compared to Posenet (a similar but lighter approach) and, with a MobileNet [39] version, OpenPose can also run on devices with low computing performance.

In videos, there are a lot of occlusions or the robot looks away from the child since its performed gesture movements. Moreover, children are always on the move and they can have difficulties with other devices, like other cameras or a Microsoft Kinect, in the therapeutic environment. With the use of high-resolution cameras or a Microsoft Kinect, we can increase the performance, limiting the portability of the system [12]. In [34], OpenPose [28] was compared with the Microsoft Kinect. The final results showed that OpenPose is accurate at recognizing gestures and it can overcome the failures of the Kinect. We assumed that the OpenPose solution on the 2D video is robust enough for gesture recognition. In a preliminary analysis, we found that it is much more accurate than Kinect when there are occlusions in the videos.

A secondary aim of our investigation was to evaluate alternative technologies to allow recognition in real-time, which may prompt autonomous adjustments of the robot's behaviour to the child's performance level during the therapy. To this end, we tested the inference time of our recognition system on an NVIDIA Jetson TX2 to explore the performance for a possible real-time gesture recognition when the robot has AI acceleration integrated onboard.

We would specify that the focus on this real-world application and low resources makes unfeasible the use of large deep neural network architectures, which would require significant additional computation on the robot and drain the battery very quickly. We could also consider the use of cloud computing, but this is not simply applicable to this clinical application, indeed streaming clinical therapy sessions over the network will create significant security concerns due to the sensible and private nature of the data and, therefore, sig-

Fig. 1 In these examples you can see that the children's height is always smaller than the caregivers height

nificant overheads and further delays to secure the data via encryption/decryption.

3.1 Gesture recognition approach

The proposed method is divided into three steps. In the first step, each frame is computed by OpenPose [28] which is a real-time pose estimator. OpenPose returns the human pose in a reasonable time which depends on the computational power, extracting the pose from the image by a deep network based on the CNNs. See for instance Fig. 4. We considered only the child's pose, discarding the caregiver's pose, which differs in greater height (see Fig. 1).

Nevertheless, OpenPose was able to extract the human joints even if they are lacking. After gathering data from each video, transformed in human pose sequences with 18 joints, we normalised data according to the following equations that are applied for each joint (X, Y) assuming that the image centre is the origin $(0, 0)$:

$$X = \lfloor X + 0.5 * width \rceil \, ; \, Y = \lfloor Y + 0.5 * height \rceil$$

where $width$ and $height$ are the image dimensions of the video. Normalisation allows gestures to be better described by making the data invariant with respect to the person's height and positioning relative to the sensor.

In the second step, the human poses extracted from each video frame are given as input to a deep model based on LSTMs like in [40]. This model (see Fig. 2) automatically extracts the temporal features of the pose sequence. We used 84 and 66 units respectively for the first and the second LSTM layer for the "Already Seen" setting while 80 and 64 units for the "Leave Child Out" and "Interleave" settings. The number of epochs was 300 to train the different models. The kernel initializer was the *Xavier uniform* and the optimisation algorithm for gradient descent was *Adam*.

The final step consists in classifying the gestures by a full-connected layer. As an activation function we used softmax and the number of nodes of the full-connected layer is 5 corresponding to the number of classes. During the experiments, 207 videos of about 1.10 minute and about 10 fps were recorded for six children. The gestures are four: "kiss", "clap the hands", "greeting", "raise the arms". We also added a "failure" class to label imitation failures.

Then we trained our model with different configurations using a sliding window approach (see Fig. 3) with one and two steps. We can also find a step approach in [41] and in [40] where the authors combine the results with different steps, considering different temporal scales, in contrast to us

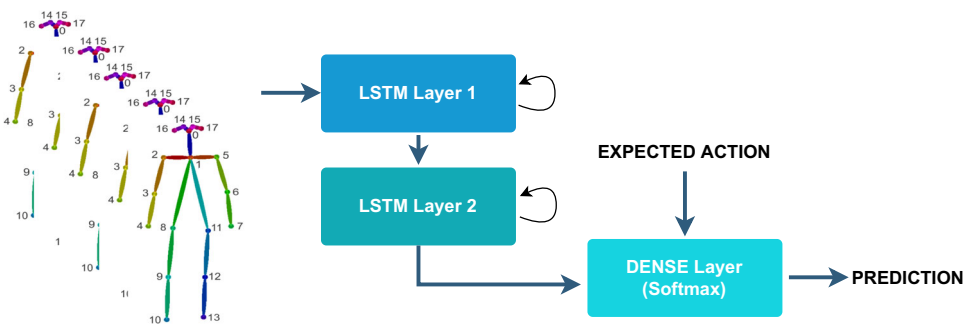


Fig. 2 Our model is based on two layers of LSTMs that take as an input the skeleton sequence and the expected action and it gives as an output the action/activity performed by the user. Remember that the goal is not the prediction of the action itself, but the verification of whether the child has imitated the robot's movements

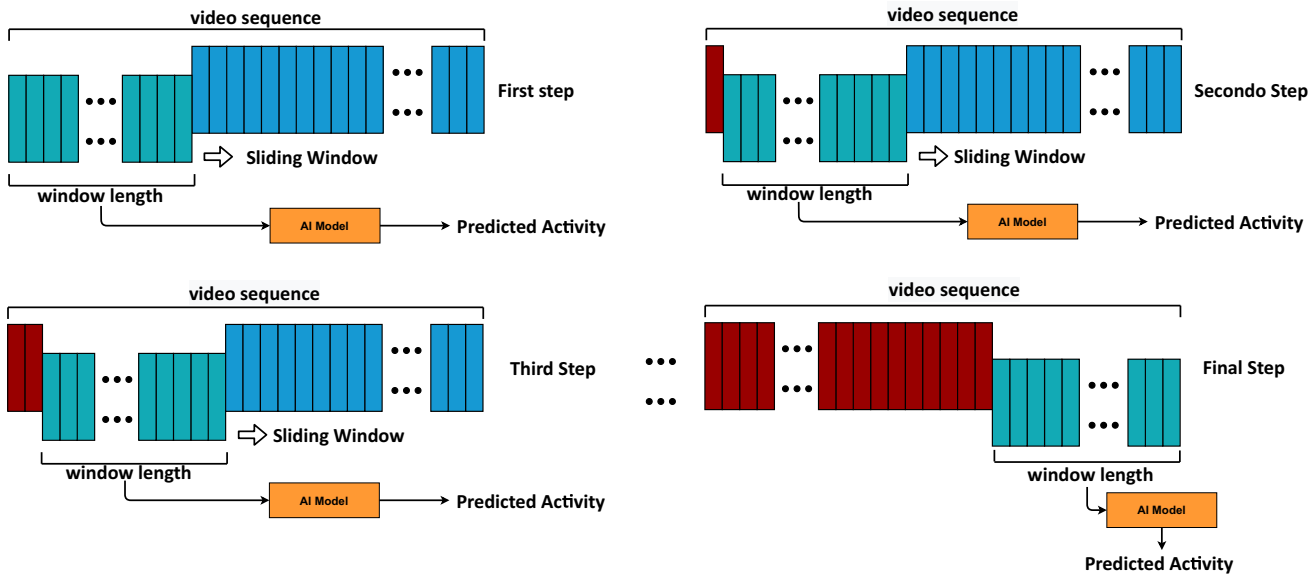
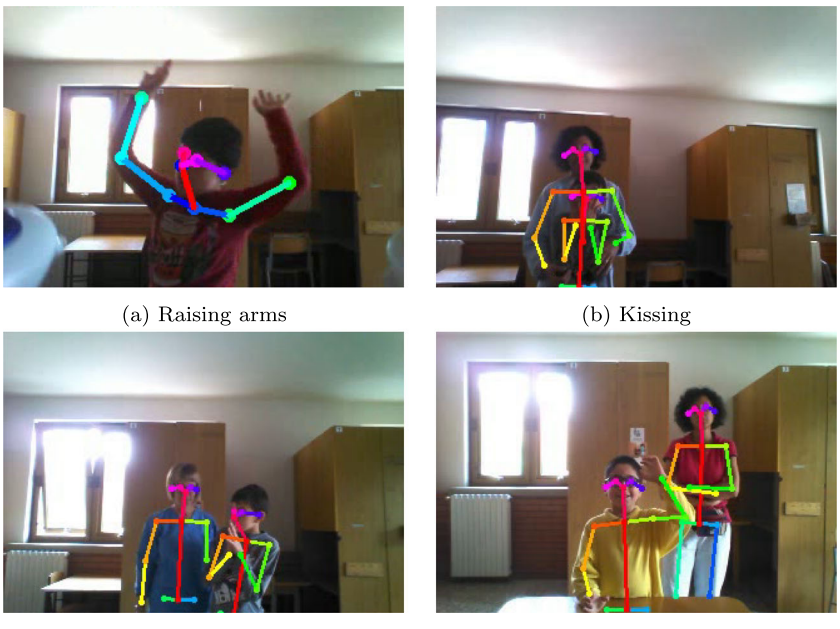


Fig. 3 A diagram graphically explaining the sliding window approach. The information from the sliding window is processed recursively along the video sequence frames and, at the end of this recursively process, this approach returns the final activity or action performed. Activity or action is predicted for each sliding window. The final activity is the one with the highest frequency

who do not combine the different steps. We used a sliding window of 5, 10, 15, 20, 25 sequence frames. The input of the model is composed of a sequence of human skeleton joints normalised according to the image dimensions and the label of the gesture performed by the robot ("kiss", "clap the hands", "greeting", "raise the arms"). The output is one of the four gesture labels or the label "failure" in case the child fails to imitate the robot.

The deep model based on LSTMs is composed of two LSTM layers that take in input the pose sequence. The features extracted from the sequence are concatenated with the gesture label ("kiss", "clap the hands", "greeting", "raise the



(a) Raising arms (b) Kissing

(c) Clapping hands (d) Greeting

Fig. 4 Four frame video showing the children (and their caregivers) with their skeleton joints recognised by OpenPose [28]

arms") that is encoded using the one-hot-encoding process that which refers to the gesture performed by the robot.

3.2 Settings

Three evaluation settings are proposed to assess the results of our approach: **Already Seen**, proposed [42] as "have seen", in which the training data is composed by five children and a half of the sixth child's data that is taken randomly; the test data is the remaining of the sixth child's data; **Leave Child Out**: the model was trained on five children and tested on the sixth; in literature, we can find the same configuration named as "new person" or "leave-one-out cross-validation" [42]; **Interleave**, similar to the "Leave Child Out" setting, but the gestures of different children were interleaved to take into account the significantly different quality and efficacy of the gesture executions.

3.3 Comparison with classical ML methods for gesture recognition

We compared the type of approach proposed with classical machine learning methods using Weka [43]. We have tested these algorithms both with and without normalisation. The results show a general improvement in accuracy without normalisation with respect to frame resolution. The pose sequences have been processed to extract the 5 most significant poses. We applied K-means, a clustering algorithm, to search for 5 clusters. Then we identified the 5 centroids that represent the 5 most significant poses that identify the sequence of the gesture. The 5 poses extracted for each instance are the samples of our ML classifier training dataset. We used the following classification algorithms [44] which are models of supervised learning to compare our proposed approach:

- Bayesian Network is a probabilistic model that represents a set of stochastic variables with their conditional dependencies using a DAG (direct acyclic graph);
- HMM (Hidden Markov Model) is a Markov chain in which states are not directly observable and is widely used in the recognition of the time pattern of time series;
- Naive Bayes is a simplified Bayesian classifier that assumes assumptions of independence of characteristics;
- SVM (Support Vector Machine) is a model that represents data as points in space, mapping them in order to define the belonging of each data to a class;
- J48 [45] is the implementation in Weka of the C4.5 algorithm, based on decision trees;
- Random Forest is a classifier obtained from the aggregation of multiple random decision trees;
- Random Tree is based on random decision trees.

4 Results

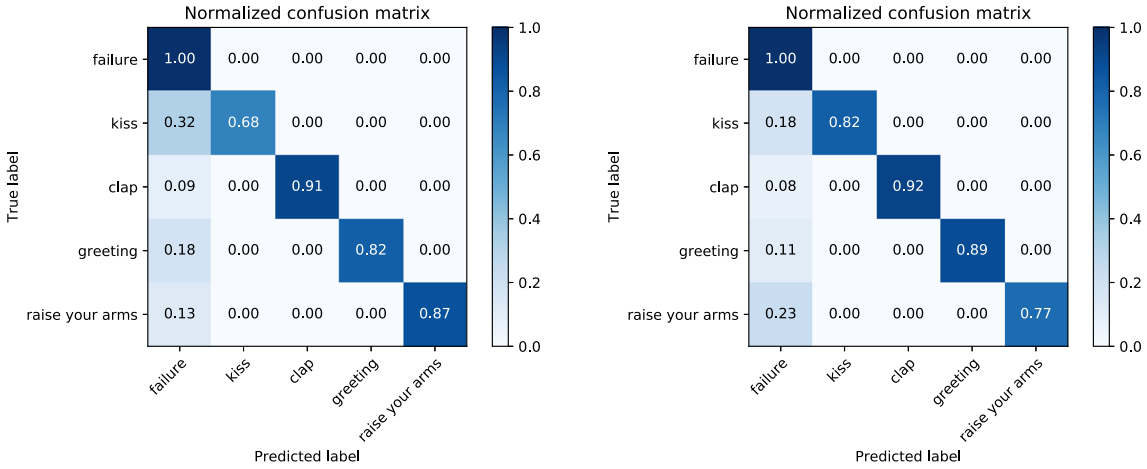
Three different settings, two different steps and five different timesteps are tested using our deep LSTM model obtaining the results shown in Table 1. Figure 5 shows two confusion matrices for the AlreadySeen setting with 1 and 2 frames. The final average accuracy result has a very good result since the number of instances of failures is almost equal to the sum of successes. In general, however, we have very good recognition of failures and successes of the children's imitation despite the NAO camera movement during gesture execution and despite the low resolution. We would like to emphasise the best results (see Tables 1 and 2) with a timestep of 5 and in general the tendency to overcome the 90.00% of accuracy. We want to underline the worst accuracy with "Interleave" and timestep 25 which is 87.13% of accuracy with step 1 and 87.06 of accuracy with step 2. The results gradually rise decreasing the timestep. Indeed, we have the best accuracy results in the setting "Already Seen" with 94.56% and 94.13% for steps 1 and 2.

4.1 Computational performance and power consumption evaluation

The execution time on 1000 frames of OpenPose takes on average 0.13 ± 0.01 sec on each frame while our model takes on average 0.03 ± 0.00 sec on an entire sequence of 25

Table 1 Accuracy results for the three settings with a step of 1 frame using our method

Timestep	Setting	Accuracy (%)	Mean (%)
5	AlreadySeen	94.37 ± 2.34	90.59 ± 7.52
5	Interleave	88.32 ± 3.86	
5	LeaveChildOut	89.08 ± 12.07	
10	AlreadySeen	76.63 ± 37.64	84.80 ± 22.07
10	Interleave	88.11 ± 3.90	
10	LeaveChildOut	89.66 ± 10.11	
15	AlreadySeen	92.35 ± 3.92	88.92 ± 7.80
15	Interleave	87.45 ± 4.49	
15	LeaveChildOut	86.95 ± 12.25	
20	AlreadySeen	93.20 ± 3.55	88.38 ± 7.15
20	Interleave	85.21 ± 3.26	
20	LeaveChildOut	86.74 ± 10.36	
25	AlreadySeen	93.45 ± 2.69	88.35 ± 8.11
25	Interleave	86.54 ± 2.83	
25	LeaveChildOut	85.06 ± 12.66	



(a) 1 frame (b) 2 frame

Fig. 5 Two confusion matrices for the setting AlreadySeen with 1 and 2 frames

frames. We used the Mobilenet network in OpenPose algorithm to decrease the computational time on the Jetson TX2. We compared our method with classical machine learning algorithms (Table 3). The classifiers used in the comparison are the following: SVM (Support Vector Machine), Bayesian Network, HMM (Hidden Markov Model), J48, Random Forest, and Random Tree. The results of the SVM and the HMM algorithms are identical while in general all the other algorithms, except the Random Forest, have statistically worse results than the SVM and HMM algorithms at a significance level of 0.05. The Random Forest algorithm performs better than the SVM and the HMM only in the "AlreadySeen" setting. In short, our deep model has statistically better results than all the tested machine learning algorithms at the significance level of 0.05. One of the additional information we have is the behaviour of the robot that the child must imitate. In the final results, we have noticed that they improve slightly by adding this information to the 5 poses extracted from the sequence of the gesture.

Table 2 Accuracy results for the three settings with a step of 2 frames using our method

Timestep	Setting	Accuracy (%)	Mean (%)
5	AlreadySeen	95.09 ± 2.48	85.12 ± 22.47
5	Interleave	88.31 ± 4.13	
5	LeaveChildOut	71.97 ± 36.79	
10	AlreadySeen	92.47 ± 3.44	89.10 ± 7.57
10	Interleave	87.50 ± 3.01	
10	LeaveChildOut	87.33 ± 12.38	
15	AlreadySeen	93.34 ± 2.61	89.16 ± 6.04
15	Interleave	86.62 ± 3.43	
15	LeaveChildOut	87.51 ± 8.58	
20	AlreadySeen	93.64 ± 2.38	89.12 ± 6.56
20	Interleave	85.98 ± 2.87	
20	LeaveChildOut	87.73 ± 9.67	
25	AlreadySeen	92.11 ± 2.94	87.78 ± 7.63
25	Interleave	86.10 ± 1.81	
25	LeaveChildOut	85.14 ± 12.32	

Finally, we performed a power consumption analysis in order to provide an indicative evaluation for the future integration of an edge AI board like the NVIDIA Jetson TX2 into the robot. The analysis was made by measuring the current drawn by the board and the supply voltage from the standard power brick (AC to DC power converter). First, we measured the baseline current, which was on average $240mA$, with 20 a standard deviation (st. dev.), then the current drawn during the inference, which was on average $491mA$ with $38mA$ st. dev. and a peak of $533mA$. The supply voltage was almost constant at $19V$ with 0.02 st. dev. This result shows that the gesture recognition with our method consumes on average only $4.77W$ on the NVIDIA Jetson TX2. The peak consumption is $10.14W$ (including the baseline consumption).

This power consumption is theoretically compatible with the battery specifications of a small robot like NAO, which has a $48.6Wh$ battery with a nominal voltage of $21.6V$ and a maximum current of $2A$, with a maximum peak consumption of $43.2W$.

5 Discussion

The results present a solution posed by the clinical requirement to not introduce other devices, indeed the only device used to acquire data was the built-in camera of the NAO

Table 3 Accuracy results for the three settings with ML methods

Classifier	Setting	Accuracy (%)	Mean (%)
SVM	AlreadySeen	66.38	65.67
	Interleave	63.64	
	LeaveChildOut	66.99	
Bayesian Network	AlreadySeen	33.90	33.34
	Interleave	32.55	
	LeaveChildOut	33.56	
HMM	AlreadySeen	66.38	65.67
	Interleave	63.64	
	LeaveChildOut	66.99	
Naive Bayes	AlreadySeen	29.08	27.12
	Interleave	26.29	
	LeaveChildOut	25.98	
J48	AlreadySeen	62.84	58.53
	Interleave	56.89	
	LeaveChildOut	55.85	
Random Forest	AlreadySeen	72.03	66.24
	Interleave	62.15	
	LeaveChildOut	64.55	
Random Tree	AlreadySeen	61.41	53.96
	Interleave	48.80	
	LeaveChildOut	51.67	

robot, which operates at a low resolution (320 x 240) and a low frame rate (10 fps).

Our method incorporates a variety of techniques including motion capture, computer vision, and machine learning to accurately recognize the gestures of autistic children. By considering the entire body, we are able to capture a wider range of subtle movements that may be missed by methods that only focus on the hand.

The results show that the proposed algorithm was able to efficiently deal with the lack of depth information by extracting the 2D poses of the children with the OpenPose algorithm.

Another practical problem was the motion of the NAO while performing gestures. The video recorded by the camera fixed on the robot's forehead was unstable and, in a few cases, the vision of the child's movements was partially occluded because of the robot's movements (head, torso, arms and hands). The solution to this problem was investigated using different timesteps and taking the most likely results from a time window that corresponds to the time spent by the robot performing the gesture to imitate.

Another issue faced is that the dataset is unbalanced since it has multiple instances of children's failures: the sum of gestures on the test set is about half of the number of failures. Although the results of the LSTM model with 1 step and 5 timestep are slightly better, in general, the 2 step behaves well with the various timesteps. This result is useful for reducing the performance, indeed the 2 steps approach has a shorter inference time using an embedded AI acceleration device like the NVIDIA Jetson TX2, which mixes good performance and low power consumption. In practice, it reduces the computation almost by half by applying OpenPose every two frames.

We highlight that the LSTM model has significantly exceeded the results of the machine learning algorithms proposed for comparison. We would like to remark that articles mentioned in the related work report an accuracy of around 90-93% with synthetic data, our approach achieves the same levels of accuracy with real-world data Furthermore, by considering the entire body, we are able to provide a more comprehensive understanding of the child's behaviour and their attempts to interact with the robot.

We also provided a proof-of-concept evaluation of the use of state-of-the-art off-the-shelf embedded systems for edge-AI. We tested the performance of the NVIDIA Jetson TX2 (see Fig. 6) which is increasingly used in studies that require AI algorithms to run on low-cost, low-power platforms [46]. This proof-of-concept demonstration provides experimental information that will guide the design of future robots for robot-led therapy that will be able to provide

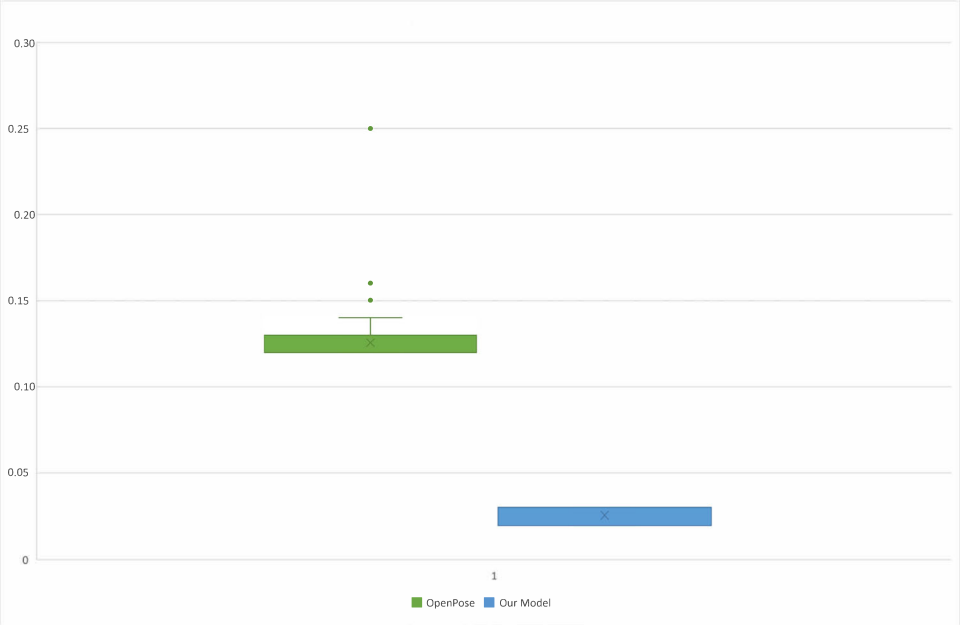


Fig. 6 Computational performance in seconds of the OpenPose algorithm and our model on NVIDIA Jetson TX2

real-time evaluation of the children's behaviour, therefore adapt the interaction to personalise the clinical intervention autonomously.

6 Conclusion

In this work, we studied the automation of imitation recognition during robot-assisted training of children with Autism Spectrum Disorder. Indeed, we used a new data set collected during a clinical study with children with ASD in a real unconstrained setting. The clinical study provided low-resolution videos recorded by the robot camera during the robot-led therapy. The aim of the automation of this task is to guarantee the objectivity of the evaluation and provide data for continuous assessment of the progress during the therapy. This technological solution can overcome the limitations of manual annotation which is a long and tedious process which requires multiple assessors to ensure impartiality, with a considerable cost for the healthcare providers. From an applied perspective, a fundamental point of our approach was to comply with the clinical requirements, i.e. to reduce the intrusion by using only the camera that is embedded in the robotic platform. Indeed, children with ASD may be upset by the introduction of many novel items in their environment, therefore, simplicity is an essential pre-requisite for the inclusion of any technology in the actual therapy. At the same time, this creates a technological challenge because the embedded camera does not provide the depth measurement and was only able to acquire images with a frequency of 10 fps and a resolution of $320x240$ pixels because of the limitations of the onboard computational resources (CPU and memory). Considering the lack of depth images, we opted for the OpenPose algorithm which is more accurate than Microsoft Kinect when there are occlusions in videos. The proposed method to automatically evaluate the gesture is a deep model based on LSTMs. Three settings were used to test the model: "AlreadySeen", "Interleave", "LeaveChildOut". To enhance the performance of the deep model, we tested five different timesteps (5, 10, 15, 20, 25) and two steps (1 and 2). The final results show a very good accuracy: on average the 93.01% of accuracy with timestep 5 and step 1. We wanted to compare these results with some classic machine-learning algorithms. The results of the deep model are statistically better than the proposed ML algorithms at the significance level of 0.05. Finally, given the low computational power of the NAO robot, in order to evaluate the performance level of imitation training during the therapy, we tested our model with OpenPose on an NVIDIA Jetson TX2, which is an embedded AI computing device. In the production stage, we can say that the deep LSTM model with step 2 would reduce by half the computational time to predict the gesture for the calculation of joints with OpenPose. In short, the calculation of joints with a step equal to 2 is not done for each frame but for every two frames.

Funding This work was supported by the European Commission Horizon 2020, grant numbers: 955778 (PERSEO), 703489 (CARER-AID). The work of Alessandro Di Nuovo acknowledges the support of the UK Engineering and Physical Sciences Research Council (EPSRC), grant numbers: EP/X018733/1 (ALDENS), EP/P030033/1 (NUMBERS). The work of Silvia Rossi has been partially supported by the Ital-

ian Ministry for Universities and Research (MUR) under the grants FIT4MEDROB (MUR: PNC0000007) and FAIR (MUR: PE0000013). Author Daniela Conti acknowledges the support of the University of Catania PIACERI Starting Grant Line 3 for the "START to Aid" project (209564).

Data Statement The raw data that support the findings of this study are not openly available due to parents' requests and for the privacy of patients and therapists involved. It will be made available from the corresponding author upon reasonable request via email.

Declarations

Compliance with Ethical Standards Ethical approval for the clinical study and the use of the data was obtained from both the ethical council of IRCSS Oasi Maria SS of Troina, where the patients were hospitalised, and Sheffield Hallam University, which was the leading University for this study. All the parents signed consent forms that allowed us to collect data during the study and use it only for research purposes.

Conflict of interest The authors declare no conflict of interest for the study reported in this article.

References

1. Provoost S, Lau HM, Ruwaard J, Riper H (2017) Embodied conversational agents in clinical psychology: a scoping review. J Med Internet Res 19:e151
2. Scoglio AA, Reilly ED, Gorman JA, Drebing CE (2019) Use of social robots in mental health and well-being research: systematic review. J Med Internet Res 21:e13322
3. Pandey AK, Gelin R (2018) A Mass-Produced Sociable Humanoid Robot: Pepper: The First Machine of Its Kind. IEEE Robot & Autom Mag 25:40–48. https://doi.org/10.1109/MRA.2018.2833157
4. Belpaeme T, Kennedy J, Ramachandran A, Scassellati B, Tanaka F (2018) Social robots for education: A review. Sci Robot 3:eaat5954. https://doi.org/10.1126/scirobotics.aat5954
5. Matarić MJ, Scassellati B (2016) Socially Assistive Robotics. In: Siciliano B, Khatib O (eds.) Springer Handbook of Robotics. Springer International Publishing, Cham, pp 1973–1994. https://doi.org/10.1007/978-3-319-32552-1_73
6. Di Nuovo A, Broz F, Wang N, Belpaeme T, Cangelosi A, Jones R, Esposito R, Cavallo F, Dario P (2018) The multi-modal interface of Robot-Era multi-robot services tailored for the elderly. Intell Serv Robot 11:109–126. https://doi.org/10.1007/s11370-017-0237-6
7. Loomes R, Hull L, Mandy WPL (2017) What is the male-to-female ratio in autism spectrum disorder? a systematic review and meta-analysis. J Am Acad Child Adolesc Psychiatry 56:466–474
8. Conti D, Cirasa C, Di Nuovo S, Di Nuovo A (2020) Robot, tell me a tale!: A Social Robot as tool for Teachers in Kindergarten. Interact Stud 21:220–242
9. Scassellati B, Admoni H, Matarić M (2012) Robots for Use in Autism Research. Annu Rev Biomed Eng 14:275–294. https://doi.org/10.1146/annurev-bioeng-071811-150036
10. Wood LJ, Zaraki A, Robins B, Dautenhahn K (2019) Developing Kaspar: A Humanoid Robot for Children with Autism. Int J Soc Robot. https://doi.org/10.1007/s12369-019-00563-6
11. Cao H, Esteban PG, Bartlett M, Baxter P, Belpaeme T, Billing E, Cai H, Coeckelbergh M, Costescu C, David D, Beir AD, Hernandez D, Kennedy J, Liu H, Matu S, Mazel A, Pandey A, Richardson K, Senft E, Thill S, Perre GVd, Vanderborght B, Vernon D, Wakanuma K, Yu H, Zhou X, Ziemke T (2019) Robot-Enhanced Therapy: Development and Validation of Supervised Autonomous Robotic System for Autism Spectrum Disorders Therapy. lIEEE Robot Autom Mag 26:49–58. https://doi.org/10.1109/MRA.2019.2904121
12. Di Nuovo A, Conti D, Trubia G, Buono S, Di Nuovo S (2018) Deep learning systems for estimating visual attention in robot-assisted therapy of children with autism and intellectual disability. Robotics 7:25
13. Boucher J (1977) Alternation and sequencing behaviour, and response to novelty in autistic children. J Child Psychol Psychiatry 18:67–72
14. Conti D, Trubia G, Buono S, Di Nuovo S, Di Nuovo A (2021) An empirical study on integrating a small humanoid robot to support the therapy of children with autism spectrum disorder and intellectual disability. Interact Stud 22:177–211
15. Gouaillier D, Hugel V, Blazevic P, Kilner C, Monceaux J, Lafourcade P, Marnier B, Serre J, Maisonnier B (2009) Mechatronic design of NAO humanoid. 2009 IEEE International conference on robotics and automation
16. Robaczewski A, Bouchard J, Bouchard K, Gaboury S (2021) Socially assistive robots: The specific case of the nao. Int J Soc Robot 13:795–831
17. Alabdulkareem A, Alhakbani N, Al-Nafjan A (2022) A systematic review of research on robot-assisted therapy for children with autism. Sensors 22. https://www.mdpi.com/1424-8220/22/3/944. https://doi.org/10.3390/s22030944
18. Conti D, Di Nuovo S, Di Nuovo A (2021) A brief review of robotics technologies to support social interventions for older users. Human Centred Intell Syst pp 221–232
19. Schneider P, Memmesheimer R, Kramer I, Paulus D (2019) Gesture recognition in rgb videos using human body keypoints and dynamic time warping. In: Robot World Cup, Springer, pp 281–293
20. Mazhar O, Ramdani S, Navarro B, Passama R, Cherubini A (2018) Towards real-time physical human-robot interaction using skeleton information and hand gestures. In: 2018 IEEE/RSJ International conference on intelligent robots and systems (IROS), IEEE, pp 1–6
21. Ojeda-Castelo JJ, Capobianco-Uriarte MdLM, Piedra-Fernandez JA, Ayala R (2022) A survey on intelligent gesture recognition techniques. IEEE Access 10:87135–87156. https://doi.org/10.1109/ACCESS.2022.3199358
22. Qiao S, Wang Y, Li J (2017) Real-time human gesture grading based on openpose. In: 2017 10th International congress on image and signal processing, biomedical engineering and informatics (CISP-BMEI), IEEE, pp 1–6
23. Biswas KK, Basu SK (2011) Gesture recognition using microsoft kinect®. In: The 5th International conference on automation, robotics and applications, IEEE
24. Venkatnarayan RH, Page G, Shahzad M (2018) Multi-user gesture recognition using wifi. In: Proceedings of the 16th annual interna-

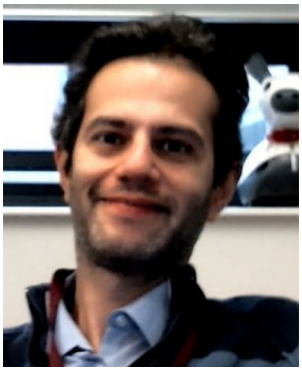
tional conference on mobile systems, applications, and services, ACM, pp 401–413
25. Zheng Y, Zhang Y, Qian K, Zhang G, Liu Y, Wu C, Yang Z (2019) Zero-effort cross-domain gesture recognition with wi-fi. In: Proceedings of the 17th annual international conference on mobile systems, applications, and services, ACM, pp 313–325
26. Li C, Zhang X, Liao L, Jin L, Yang W (2019) Skeleton-based gesture recognition using several fully connected layers with path signature features and temporal transformer module. In: Proceedings of the AAAI conference on artificial intelligence, vol 33 pp 8585–8593
27. Mazzia V, Angarano S, Salvetti F, Angelini F, Chiaberge M (2022) Action transformer: A self-attention model for short-time pose-based human action recognition. Pattern Recog 124:108487
28. Cao Z, Simon T, Wei S-E, Sheikh Y (2017) Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7291–7299
29. Liu Z, Zhang H, Chen Z, Wang Z, Ouyang W (2020) Disentangling and unifying graph convolutions for skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 143–152
30. Plizzari C, Cannici M, Matteucci M (2021) Skeleton-based action recognition via spatial and temporal transformer networks. Comput Vis Image Underst 208:103219
31. Hampiholi B, Jarvers C, Mader W, Neumann H (2023) Convolutional transformer fusion blocks for multi-modal gesture recognition. IEEE Access 11:34094–34103
32. Miki D, Kamitsuma K, Matsunaga T (2023) Spike representation of depth image sequences and its application to hand gesture recognition with spiking neural network. SIViP pp 1–9
33. Ramalingam B, Angappan G (2023) A deep hybrid model for human-computer interaction using dynamic hand gesture recognition. Comput Assist Methods Eng Sci
34. Rahman A, Clift LG, Clark AF (2019) Comparing gestural interfaces using kinect and openpose. In: CGVC, pp 103–104
35. Fang H-S, Xie S, Tai Y-W, Lu C (2017) Rmpe: Regional multi-person pose estimation. In: 2017 IEEE International conference on computer vision (ICCV), pages 2353–2362
36. Li J, Wang C, Zhu H, Mao Y, Fang H-S, Lu C (2019) Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10863–10872
37. Xiu Y, Li J, Wang H, Fang Y, Lu C (2018) Pose flow: Efficient online pose tracking. In: British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, BMVA Press, p 53. Accessed 3–6 Sept 2018
38. Bharati P, Pramanik A (2020) Deep learning techniques—r-cnn to mask r-cnn: a survey. In: Computational intelligence in pattern recognition, Springer, pp 657–668
39. Sinha D, El-Sharkawy M (2019) Thin mobilenet: An enhanced mobilenet architecture. In: 2019 IEEE 10th Annual ubiquitous computing, electronics & mobile communication conference (UEMCON), IEEE, pp 0280–0285
40. Ercolano G, Riccio D, Rossi S (2017) Two deep approaches for adl recognition: A multi-scale lstm and a cnn-lstm with a 3d matrix skeleton representation. In: 2017 26th IEEE International symposium on robot and human interactive communication (RO-MAN), IEEE, pp 877–882
41. Neverova N, Wolf C, Taylor GW, Nebout F (2014) Multi-scale deep learning for gesture detection and localization. In: European conference on computer vision, Springer, pp 474–490
42. Sung J, Ponce C, Selman B, Saxena A (2012) Unstructured human activity detection from rgbd images. In: 2012 IEEE International conference on robotics and automation, IEEE, pp 842–849
43. Desai A, Sunil R (2012) Analysis of machine learning algorithms using weka. Int J Comput Appl 975:8887
44. Alpaydin E (2014) Introduction to Machine Learning. Adapt Comput Mach Learn (3rd edn.) publisherMIT Press, Cambridge, MA
45. Mathuria M (2013) Decision tree analysis on j48 algorithm for data mining. Int J Adv Res Comput Sci Softw Eng vol 3
46. Mittal S (2019) A survey on optimized implementation of deep learning models on the nvidia jetson platform. J Syst Archit 97:428–442. https://doi.org/10.1016/j.sysarc.2019.01.011

Giovanni Ercolano is a graduate student in Computer Science with a master's degree and a PhD in Computer Science and Electrical Engineering (ITEE) from the University of Naples Federico II. During my academic career, I focused on human-robot interaction and the development of deep learning algorithms for human activity classification. In particular, I contributed to the User-centered Profiling and Adoption for Socially Assistive Robotics (UPA4SAR) project, where I participated in the implementation and real-world testing of a socially interactive robot designed to autonomously assist the elderly within their homes.

Silvia Rossi is an associate professor at the Department of Electrical Engineering and Information Technologies, University of Naples Federico II, where she is the scientific director of the PRISCA Lab (Projects of Intelligent Robotics and Advanced Cognitive Systems). She received the M.Sc. degree in Physics from the University of Naples Federico II, Italy, in 2001, and the Ph.D. in Information and Communication Technologies from the University of Trento, Italy, in 2006. Prof. Rossi has been involved in several EU and non-EU projects. She is currently the principal investigator and coordinator of the MSCA-ITN-2020 PERSEO (European Training Network on Personalized Robotics as Service Oriented applications), PI of the HORIZON-TMA-MSCA-DN project TRAIL (TRAnsparent, InterpretabLe Robots), and Coordinator of the national PRIN project ADVISOR (ADaptiVe legIble robotS for trustwORthy health coaching). She was the general chair of RO-MAN 2020 and RO-MAN 2022 and she is in the program committee of several international conferences on human–robot interaction and artificial intelligence. Her research interests include Socially Assistive Robotics, Human-Robot Interaction, Cognitive Architectures, and User Profiling and Recommender Systems. Her main research activities aim at the investigation of computational approaches for autonomous agents' behaviors able to interact and support people by extracting meaningful information to model the user and to adapt the agent behavior. She published more than 180 papers in international journals, books, and conferences.

Daniela Conti is currently an Assistant Professor (Tenure Track) in the Department of Humanities at the University of Catania. She is a graduate (B.Sc. and M.Sc.) in Psychology (2008, 2010) and B.Sc. in Psychiatric Rehabilitation and Social Education (2002), all awarded with the highest distinction (110/110 cum laude), and received the PhD in Neuroscience at the University of Catania, Italy (2016). Her work mainly focuses on Artificial Intelligence, the applicability of robotics to autism spectrum disorder with intellectual disability, and the acceptability of robotics in clinical and educational settings.

Author of several scientific publications, her work has been supported by the H2020 research and innovation program of the European Union, CARER-AID, project "Controlled Autonomous Robot for Early Detection and Rehabilitation of Autism and Intellectual Disability", Marie Sklodowska Curie Individual Fellowship in UK. Since 2021 she is a member of the Editorial Board of the international journal "Interaction Studies". Member of the Italian Association of Psychology - Experimental Psychology section, since 2020. Member of the European Network for the Advancement of Artificial Cognitive Systems, Interaction and Robotics (EUCOG), since 2014. She is a licensed clinical psychologist certified by the National Board of Psychologists (Italy), since September 2011 (A-6007).



Alessandro Di Nuovo is Professor of Machine Intelligence at the Department of Computing, Sheffield Hallam University (SHU). He received the Laurea (M.Sc.Eng.) and Ph.D. degrees in Informatics Engineering from the University of Catania, Italy, in 2005 and 2009, respectively. From 2012 to 2015, he was a Research Fellow with the University of Plymouth, U.K.

Prof. Di nuovo is the leader of AI, Robotics and Digital for the SHU Advanced Wellbeing Research Institute. He is the founder and leader of the Smart Interactive Technologies (SIT) Research Laboratory, which has cutting-edge facilities and equipment for conducting internationally renowned research in interdisciplinary applications of machine intelligence, including healthcare and well-being. Currently, he is the scientific coordinator of the Horizon Europe project "Performance in Robots Interaction via Mental Imagery" (PRIMI), which was awarded €7.3 million for 50 months, from 2023–2027.

Since 2021, I am serving as Topic Editor-in-Chief of the International Journal of Advanced Robotic Systems (Sage). I am also serving as Associate Editor for the IEEE Journal of Translational Engineering in Health & Medicine, Applied Sciences and Robotics journals (MDPI). For his academic and professional service, in 2014, he was awarded the status of Senior Member of the IEEE.