Predicting apparent personality from body language: benchmarking deep learning architectures for adaptive social human–robot interaction

Marta Romeo, Daniel Hernández García, Ting Han, Angelo Cangelosi and Kristiina Jokinen

ABSTRACT
First impressions of personality traits can be inferred by non-verbal behaviours such as head pose, body postures, and hand gestures. Enabling social robots to infer the apparent personalities of their users based on such non-verbal cues will allow robots to gain the ability of adapting to their users, constituting a further step towards the personalisation of human–robot interactions. Deep learning architectures such as residual networks, 3D convolutional networks, and long-short time memory networks have been applied to classify human activities and actions in computer vision tasks. These same architectures are beginning to be applied to study human emotions and personality by focusing mainly on facial features in video recordings. In this work, we exploit body language cues to predict apparent personality traits for human–robot interactions. We customised four state-of-the-art neural network architectures to the task, and benchmarked them on a dataset of short side-view videos of dyadic interactions. Our results show the potential for deep learning architectures to predict apparent personality traits from body language cues. While the performance varied between models and personality traits, our results show that these models could still be able to predict sole personality traits, as exemplified by the results on the conscientiousness trait.

1. Introduction
Personality computing is considered fundamental for a variety of life aspects: from one’s own psychological well-being to occupational and relational choices [1]. It aims to solve three main problems: automatic personality recognition (the recognition of the true personality of an individual), automatic personality perception (the prediction of the personality others attribute to a given individual, the apparent personality), and automatic personality synthesis (the generation of artificial personalities through embodied agents) [1]. In this paper, we focus on personality perception, otherwise known as automatic Apparent Personality Prediction (APP).

Studies such as [2–4] have shown that human–robot interactions depend on personality computing as much as human–human interactions. Hence, if social robots were able to detect the personality of their users, they would be equipped with an important tool that they could use to adapt to the people they are interacting with, consequently improving the quality of the interaction and their users’ general well-being [5].

Body language cues (e.g. head pose, gaze, facial expressions and body language) provide fundamental information to form a first impression of the personality traits of a person [6]. Observing facial features and eye contact duration [7], as well as the frequency and amplitudes of the head movements, hand gestures and shifts in body postures, facilitates humans to infer certain personality traits of their interlocutors [8]. Likewise, a robot equipped with the ability of detecting and analysing such non-verbal behaviours should also be able to infer the apparent personality of its interlocutors, creating the first impression even before the actual interaction starts. By observing people interacting with each other, robots may learn to predict the apparent personality of their potential users and use this information at the initiation of contact, to make their users more at ease when directly addressing them [4].

In this work, we aim to answer the following research question: what algorithm is best to equip robots with this ability, so that they can approach their users and start an interaction in the most suitable way?
We hypothesise that Deep Learning (DL) architectures are effective for the task, as they have been successfully used to identify human activities from images and video recordings [9] and also start to attract attention from the human–robot interaction community in the field of emotion recognition [10] and personality detection [11].

In fact, from [11, 12] it is clear that personality computing, and especially APP benefits from the advancements in DL. However, not many of the proposed solutions for APP take into consideration bodily signals, even though they could greatly benefit apparent personality trait analysis [13]. Most of the works taking a closer look at bodily social signals are mainly based on hand-crafted features and standard machine learning approaches in multimodal settings, like [14, 15]. In both works, high-level features including body motion, body activity and gaze are extracted from the input data and used to classify the apparent personality traits through regression and Support Vector Machines (SVM).

The state-of-the-art performance of DL for APP has been attained on datasets not taking into consideration bodily signals, but rather focusing on facial features [16, 17]. For instance, the ChaLearn Looking at People Apparent Personality Analysis competition proposes a challenge on first impressions recognition aiming to recognise apparent personality traits from short videos of people directly looking at the camera. The outcomes of this challenge indicate that DL models can be effective for predicting apparent personality from video input data. The first [18], the second [19], and the third-place [20] winners of the competition use both audio and visual information available in the ChaLearn dataset [16]. In addition, 3DCNNs have also been successfully applied to human action and activity recognition tasks [21].

Motivated by the above successes, we take a step further to investigate whether these state-of-the-art DL architectures are also effective at predicting the first impressions of the BIG5 personality traits [22] given body language data of individuals from side-view videos. To this end, we adapt a dataset that provides side-view RGB video recordings of 12 interaction sessions, with each session involving two participants (for a total of 18 participants, 9 women and 9 men). In the interactions, participants are seated face to face and take turns asking a set of questions to each other. Each interaction lasts 10–15 min, resulting in 290 short clips of Human–Human Interactions (HHI).

We decided to use the recordings of HHI from the MHHRI dataset as input for our DL architectures for the following reasons:

1. The HHI recordings provide unobstructed third-view data, captured through one static RGB Kinect sensor, which are considered most suitable for predicting apparent personalities from body language.
2. The process from which the annotations of the personalities are derived is explained and considered reliable in dictating the ground truth labels.

2. Dataset and pre-processing pipeline

Very few datasets exist for vision-based APP and almost all of them comprise videos, or still images, recorded from frontal-view cameras [12]. The ChaLearn First Impression dataset [16] provides both audio and visual information from 10,000 clips (average duration 15 s) extracted from more than 3000 different YouTube videos of people facing and speaking into the camera. The VLOG dataset [17] consists of videos between 1 and 6 min long from 469 different vloggers. The ELEA corpus [23], gathered with the aim of analysing emergent leadership in newly formed groups, collects data from 40 meetings composed of 3 or 4 members and annotated with self-reported and perceived personality.

The Multimodal Human–Human–Robot Interaction (MHHRI) dataset [24], in contrast, provides visual data from full-body movements rather than just head and facial features. Although the dataset of [23] also contains side-views captures, they are of four people sitting around a table, which result in partially occluded data. This is why we adapt the MHHRI dataset for our task.

2.1. The multimodal human–human–robot interaction dataset

The MHHRI dataset comprises both human–human and human–human–robot interactions. In addition to first-person vision data, it also provides side-view RGB video recordings of 12 interaction sessions, with each session involving two participants (for a total of 18 participants, 9 women and 9 men). In the interactions, participants are seated face to face and take turns asking a set of questions to each other. Each interaction lasts 10–15 min, resulting in 290 short clips of Human–Human Interactions (HHI).

We decided to use the recordings of HHI from the MHHRI dataset as input for our DL architectures for the following reasons:

1. The HHI recordings provide unobstructed third-view data, captured through one static RGB Kinect sensor, which are considered most suitable for predicting apparent personalities from body language.
2. The process from which the annotations of the personalities are derived is explained and considered reliable in dictating the ground truth labels.
The MHHRI dataset provides meta-data, derived from BFI-10 questionnaires [25] filled in by participants for both self-assessment and acquaintance assessment (the assessment of the other participants partaking in the study) of the BIG5 personality traits. The BIG5 Model [22] identifies five major traits that correspond to individual differences in human behaviour, way of thinking, and feelings:

(i) extroversion (being sociable, playful, assertive, etc.),
(ii) agreeableness (being appreciative, kind, etc.),
(iii) conscientiousness (being organised, efficient, etc.),
(iv) neuroticism (being insecure, anxious, etc.), and
(v) openness to experience (being intellectual, curious, etc.).

Unlike other trait theories that sort traits of individuals into binary categories, the BIG5 asserts that each personality trait is a spectrum, and each trait concurs equally to the definition of the personality of a person. Therefore, individuals are ranked on a scale between two extreme ends for each of the BIG5 traits.

To prepare the video clips derived from the HHI recordings to serve as input for the DL architectures, we pre-processed the dataset with five steps, outlined hereafter.

2.2. Cleaning the dataset

To create a consistent pool of inputs for the DL models, some of the original frames had to be discarded because they were either corrupted or they were portraying the participants with their back directed towards the camera. After this operation, over 140,000 PNG frames were left, from the 290 original HHI clips, divided among the 12 interaction sessions of the original dataset.

2.3. Creating input clips

To train the models, input clips of 16 subsequent frames were created (following the clip length used in [21]). As the original recordings of the MHHRI present a different and inconsistent number of frames, time duration, and frames per second, firstly, a step of normalising the time length of the clips was carried out to maintain temporal consistency. The original frames were grouped into 16 frames clips while trying to preserve a mean frame rate of 8 Hz and a time duration as close to 2 s as possible. This time duration was considered enough to create a first impression on the personalities of the participants depicted in the frames. This is because, according to [8, 26], an exposure time as brief as 100 ms is sufficient for individuals to form a first impression. After this step, a total of 9831 clips, with a duration between 0.96 and 4.3 s were obtained (mean = 1.88 ± 0.31 s).

2.4. Extracting individual participants

We extracted the pixels of individual participants from the frames in the dataset before feeding them into the models for training. In this pre-processing step, each frame was input to a Mask R-CNN [27] used for detecting and isolating the two participants in it. The Mask R-CNN implementation used in this work is the open-source network of [28], pre-trained on the COCO dataset [29].

The outputs of the Mask R-CNN step are the bounding boxes for the two different participants in the frame; one for the participant on the right and one for the participant on the left. The bounding boxes found by the Mask-RCNN were then used to crop the original frames of the MHHRI dataset and produce two frames, one for the right and one for the left participants. Finally, the resulting cropped images were resized to 128 × 128 pixels, while keeping the aspect ratio. An example of these steps is depicted in Figure 1.

2.5. Defining the ground truth labels

The MHHRI dataset provides self-assessment and acquaintance assessment annotations from the participants’ answers to the BFI-10 questionnaire [25]. Each of the 10 items in the BFI-10 questionnaire contributes to the score of one particular trait, with 2 BFI items for each of the BIG5 dimensions. The answers to the questionnaires are measured on a 10-point Likert scale.

We used the acquaintance assessments from the MHHRI dataset, which provide 9–12 acquaintance ratings per participant, to define the ground truths labels to train the models. A score, ranging from 1 to 10, for each participant’s ground-truth personality labels on each trait was computed by averaging over all raters. Finally, a binarisation step was carried out with respect to the mean score for each of the five traits independently, computed over all participants, to group them into two classes (i.e. low and high) per personality trait, following the procedure used in [24].

2.6. Train, test, and validation splits

In order to train and test the models, a 6-fold cross-validation was set up to evaluate their ability to learn to
Figure 1. The computation carried out by the Mask R-CNN in three stages. (a) An example of the third vision frame within the MHHRI [24] dataset given as input to the Mask R-CNN. (b) The detected masks and bounding boxes per participants in the frames. (c) The resulting left and right participants related frames, input to train the DL architectures.

| Personality trait | # clips | # participants |
|-------------------|---------|----------------|
| E (↑)             | 6160    | 8              |
| E (↓)             | 13,502  | 10             |
| A (↑)             | 10,693  | 10             |
| A (↓)             | 8969    | 8              |
| C (↑)             | 8753    | 8              |
| C (↓)             | 10,909  | 10             |
| N (↑)             | 12,123  | 9              |
| N (↓)             | 7539    | 9              |
| O (↑)             | 6661    | 10             |
| O (↓)             | 13,001  | 8              |

Personality traits are divided into two classes (low (↓) and high (↑)) per personality trait.

Table 2. The number of clips for the train, test, and validation splits of the 6-fold groups.

| Split       | G1    | G2    | G3    | G4    | G5    | G6    |
|-------------|-------|-------|-------|-------|-------|-------|
| Train=      | 29,416| 29,476| 29,590| 29,466| 29,574| 29,442|
| Validation= | 3268  | 3274  | 3286  | 3274  | 3284  | 3270  |
| Test        | 3320  | 3287  | 3224  | 3292  | 3233  | 3306  |
| Tot. clips  | 36,004| 36,037| 36,100| 36,032| 36,091| 36,018|

*Flipping data augmentation to the number of clips.

Table 1 shows the number of clips in the dataset for each personality trait and their class (high or low). The final number of clips in the \textit{train}, \textit{test}, and \textit{validation} splits in each the 6-fold cross validation groups is shown in Table 2.

3. Deep learning architectures

The architectures taking part in the ChaLearn Looking at People Apparent Personality Analysis competition are still considered the ones dictating the performance baseline. This can be seen from [11], where it is shown that DL approaches for APP builds on the winners of the ChaLearn Competition. For this reason, we chose [18–20] as the starting point for 3 of the architectures benchmarked in this work for the task of predicting apparent personality from bodily signals using side-views videos. Moreover, in [19] an additional architecture based on a 3DCNN was tested. Even though its performance were not considered satisfactory compared to the CNN + LSTM architecture that they proposed for the challenge, 3DCNNs have been successfully applied in human action and activity recognition tasks in works such as [21]. For this reason, we decided to include a 3DCNN architecture based on [21] in the pool of compared DL architectures.

The MHHRI dataset HHI was collected as a set of 12 interaction sessions, each session involving two participants. Even though some participants took part more than once, they always had a different partner to interact with. For this reason, they were considered as a separate instance of the same class. Therefore, the resulting 24 instances of person interactions (2 × 12) were grouped into six groups (G1, G2, G3, G4, G5, G6), with each group containing four instances out of the original 24. Since all sessions in the MHHRI dataset had a different number of clips, the grouping of the participants into the groups was done in a way that allowed each of the six folds to have roughly the same size in the test set (approximately 20% of the total clips).

Each model was trained six times, and each time a different group was kept out of the training set to serve as a test. This way, it was assured that the clips belonging to the test set were never seen by the models during the training phase.

After removing one of the 6 groups, creating the test set out of the input clips, a further 10% of the remaining input clips was assigned to the validation set, leaving all the remaining clips to form the training set. At this point, an additional augmentation step of mirroring each frame on their vertical axis was carried out for each of the clips in training and validation sets.

After dividing the input data into train, test, and validation sets, the clips underwent 2 final normalisation steps: \textit{mean subtraction} and \textit{rescaling}.

The final inputs to the DL architectures were the clips derived from the pre-processing pipeline described in the aforementioned five steps. Each clip was a tensor of 16 128 × 128 × 3 frames, resized and cropped with the bounding boxes obtained from the Mask R-CNN step.
The 3DCNN architecture studied in this work. The eight convolutional layers apply stride 1. The numbers in the convolutional layers represent the number of filters and the kernel size, respectively. The five max-pooling layers always apply a stride of 2 and have pool size $2 \times 2$, except for pool1, having size $1 \times 2 \times 2$ and applying stride $1 \times 2 \times 2$. In addition, not showing in the picture, a batch normalisation layer has been added after the conv1, conv2, conv3b, conv4b, and conv5b layers.

Figure 2. The 3DCNN architecture studied in this work. The eight convolutional layers apply stride 1. The numbers in the convolutional layers represent the number of filters and the kernel size, respectively. The five max-pooling layers always apply a stride of 2 and have pool size $2 \times 2$, except for pool1, having size $1 \times 2 \times 2$ and applying stride $1 \times 2 \times 2$. In addition, not showing in the picture, a batch normalisation layer has been added after the conv1, conv2, conv3b, conv4b, and conv5b layers.

The details of the four architectures benchmarked through this work are given in the following sections.

3.1. 3D deep convolutional network (3DCNN)

3DCNNs perform 3D convolutions over the spatio-temporal video volume. Unlike classical spatial 2D convolutions, 3D convolutions preserve the temporal information of the input signals. For this reason, they are better suited for learning spatio-temporal features and could be appropriate for APP from video clips.

The implemented 3DCNN in this work follows [21], and it is further described in Figure 2. An additional zero-padding operation (adding a border of pixels all with value zero around the edges of the input images) had to be carried out between conv5b and pool5, to ensure the continuity of size between the output of the convolution and the pooling layer. All convolutional layers are initialised following the Normal He initialisation [30]. Additionally, a batch normalisation layer is added before each pooling layer. All convolutional and fully connected layers, with the exception of the output layer, are activated by ReLU function.

Differently from the architecture in [21], the number of filters for each convolutional layer, and for the first two fully connected layers has been halved. This choice was made based on the necessity of reducing the number of trainable parameters to speed up the computation time while avoiding overfitting the model, too big in its original version for the amount of data available for training.

To solve the multi-label classification problem of predicting the five personality traits, the last layer in the model uses a sigmoid function for label prediction. The output of the last fully connected layer gives the probability scores for each of the five personality traits. Both the first and the second fully connected layers are followed by a dropout layer where outputs are dropped at a 50% rate.

3.2. 3D residual network (3DResNet)

The work done in [20] uses a deep residual network (ResNet) comprising of an auditory stream and a visual stream coming together in a fully connected audiovisual layer to predict personality traits from facial and auditory features.

ResNets have been successfully used in a variety of computer vision tasks [31], and the possibility of using volumetric convolutions for ResNets has been successfully explored in [32] for activity recognition from video inputs. Therefore, we expanded the network of [20] into a 3DResNet making use of volumetric convolutions. The resulting architecture is further explained in Figure 3.

The 3DResNet developed in this work has 18 layers. Each convolutional layer is initialised following the He Normal initialisation, activated by a ReLU function, and followed by a batch normalisation layer. The last fully connected layer is preceded by a global average pooling layer, and it is activated by a sigmoid function for multi-label classification.

3.3. VGG with descriptor aggregation network (VGG DAN+)

The implemented version of VGG analysed in this work is a VGG DAN+ architecture, based on the model defined in [18], which was successful in the ChaLearn 2016 competition.

The work of [18] modifies a traditional VGG-16 architecture with what is defined as a Descriptor Aggregation Network (DAN+). The main difference between the original VGG-16 architecture and the VGG DAN+ is that the last three fully connected layers of the VGG-16 are dropped and replaced by a concatenation layer, as exemplified by Figure 4.

Until the first convolution of the fifth block, the architecture follows a standard VGG-16 [33]. The difference between [33] and [18] is that, after conv5b and after pool5, a DAN+ block is added. The two DAN+ blocks
Figure 3. The architecture of the 3DResNet studied in this work. The numbers in the convolutional layers represent the number of filters, the kernel size and the stride size, respectively. Where the stride is not specified, it means that it is not applied. Similarly, for the only max-pooling layer, the first specified dimension is the pool size, and the second number refers to the stride. In addition, not showing in the picture, a batch normalisation layer has been added after every convolutional layer.

Figure 4. The VGG DAN+ architecture studied in this work. The convolutional layers do not apply any stride. In the figure, the number of filters and the kernel size are shown for each convolutional layer; the pooling layers always apply a stride of 2 and have a pool size $2 \times 2 \times 2$. In addition, not showing in the figure, a batch normalisation layer has been added after the conv1b, conv2b, conv3c, conv4c, and conv5b layers.

are then concatenated in the last step of the architecture, right before the fully connected layer. The DAN+ blocks perform a global average pooling and a global max pooling, both followed by a $L_2$ regularisation step, in parallel and on the same input they receive: the first time on the output of conv5b, and the second time on the output of pool5. The last step of the architecture is the concatenation of 4 outputs, 2 coming from the first DAN+ block and 2 coming from the second DAN+ block. Hence, the concatenation layer concatenates 2 global average pooling outputs and 2 global max pooling outputs, respectively. The last fully connected layer gives as output the probabilities score for the 5 personality traits in the clip and it is activated by a sigmoid function for multi-label classification.

The network implemented in this work differs from [18] by the following modifications: the use of volumetric 3D convolutions instead of classic 2D convolutions; the use of the Normal He kernel initialisation and the ReLU activation function for the convolutional layers; the addition of a batch normalisation layer for each of the 5 convolutional blocks, similarly to what has been done for the 3DCNN. Moreover, an additional zero-padding operation had to be carried out between conv5c and pool5, to ensure the continuity of size between the output of the convolution and the pooling layer.

3.4. CNN + LSTM network (CNN + LSTM)

Another approach to overcome the inability of 2D convolutions to capture temporal information is to combine 2D convolutional layers with recursive layers, used to learn the temporal patterns of the input [34]. This idea contrasts with the one of employing volumetric convolutions, as explored in the previous three architectures.

The use of an architecture concatenating CNN layers with a final LSTM layer has been explored before for action recognition in videos [35], and it is further
explored in [19] for learning first impressions of personality. The architecture from [19], exemplified in Figure 5, was implemented and tested as the final model under examination in this work.

Since this architecture uses 2D convolutions, instead of volumetric convolutions, in each convolutional layer, the input is only one of the 16 frames of the input clip. For this reason, the convolutions, the pooling operations, and the first fully connected layer operations are carried out in parallel 16 times, one per frame composing the clip. The 16 frames are then analysed as a sequence by the LSTM layer and by the final fully connected layer. All convolutional layers, and the first fully connected layers, are activated by a ReLU function, while the last fully connected layer is activated by a sigmoid function. Each convolutional layer is initialised following the Normal He initialisation and is followed by a batch normalisation layer. After the first fully connected and the LSTM layers a dropout layer, where outputs are dropped with an 80% rate, is added.

4. Experimental evaluation

The DL models have been trained end-to-end following a 6-fold cross-validation method, keeping as a test set one of the six groups per time, and using the train sets generated by the pre-processing pipeline of Section 2. The clips in the dataset were fed to the networks in mini batches of 12. All the models were trained using the Stochastic Gradient Descent (SGD) optimiser with momentum 0.9. The loss function SGD was optimising was the binary cross entropy. The early stopping method, monitoring the loss in the validation set, was employed to terminate the training whenever there was no substantial improvement.

To determine the best learning rate to use with each of the models, a Learning Rate Finder (LRF) was implemented. The LRF is a technique introduced in [36] that allows to identify, with a few iterations of training, a range of learning rates that would be optimal for a given model, taking into consideration the input dataset. The models were implemented using the TensorFlow open-source platform with the Keras API. The training was carried out using an Nvidia GeForce RTX 2080 GPU with 8GB of RAM. Table 3 reports the learning rate, the number of trainable parameters, the number of epochs, the time needed to train one epoch, and the final training/validation loss and accuracy for each model averaged over the six groups. All models were trained on 200 epochs but, thanks to the early stopping technique, some groups needed a lower number of epochs to reach a satisfying level of performance on the training set.

4.1. Evaluation metrics

Given the multi-label nature of the problem investigated in this work, the classical definition of metrics for binary classification would not be adequate to understand the performance of the architectures taken into consideration.

Popular metrics used for multi-label classification are the Hamming loss, Hamming score, precision, recall, $F_1$ score, subset accuracy (exact match), and subset zero-one loss.

The Hamming loss provides the proportion of labels predicted correctly [37]. The Hamming distance between two strings of equal length measures the number of positions at which the corresponding symbols are different. The accuracy taken into consideration in this work refers to the Hamming score (defined as $1 - \text{Hamming loss}$) which symmetrically measures how close the predictions are to the ground truth labels.

On the other hand, the subset zero-one loss is a generalisation of the well-known zero-one loss to the multi-label setting [38]. It requires, for each sample, that the predicted set of labels are correctly predicted as an exact match of the true set of labels. The subset accuracy (defined as $1 - \text{subset zero − one loss}$) provides the proportion of correctly classified examples [37]. The subset accuracy is a very strict evaluation measure, compared to the Hamming score, especially when the size of the label space is large.

Additionally, a per class (binary) analysis of the performance of the models for each of the five personality traits was performed. For this, precision, recall, and $F_1$ score were taken into consideration [39] together with the balanced accuracy [40]. Balanced accuracy is defined as the average of recall obtained on each class, and it is used in multi class and binary classification problems when the dataset used is unbalanced. Since the problem faced in this work is a binary multi-label problem on an unbalanced dataset (see Table 1), the balanced accuracy is used as a metric instead of the classical accuracy when analysing the performance of each trait separately.
Another useful tool used to visualise and evaluate the performance of a classifier is the Receiver Operating Characteristic (ROC) curve [41]. In the ROC curves, the true positive rate is on the $Y$-axis, and the false positive rate is on the $X$-axis. Usually, a larger area under the curve (AUC) is the sign of better output quality.

4.2. Results

A summary of the performance of the models is given in Tables 4 and 5. These results show that the 3DCNN was most successful when identifying conscientiousness; the 3DResNet when identifying conscientiousness and neuroticism; the VGGDAN+ when identifying conscientiousness and extroversion; the CNN + LSTM when identifying agreeableness.

Overall, conscientiousness is the trait for which the prediction has been mostly successful. This is also exemplified by the ROC curves for each of the BIG5 personality traits shown in Figure 6. Looking at these results for each personality trait and each architecture, summarised in Tables 4 and 5, it can be seen that not only conscientiousness was the most successful trait but also that the VGG DAN+ was the most successful architecture overall.

5. Discussion

One of the main obstacles of predicting personality traits with DL models faced by this work was obtaining appropriate training data. As previously outlined in [12], there is a lack of unified public datasets and tools to model and evaluate methodologies for APP. The MHHRI dataset used in this work was considered the most suitable for predicting the first impressions of personalities from non-verbal bodily social signals from side-view videos.

Even if the MHHRI dataset provides structured and complete data, there are some limitations that hinder the modelling task. First of all, the structure of the data is inconsistent across the HHI sessions in terms of the number of frames, session duration, and frame rate. Therefore, we carried out a thorough pre-processing to structure the data into a consistent format. Despite this, the dataset remained unbalanced (shown in Table 1) in terms of the representation of the 'high' and 'low' classes for each trait, leading to an additional unbalanced division of the available labels in different groups. Moreover, the visual data in the MHHRI dataset recordings are less informative than ideal: participants were too few and they were always portrayed in each frame sitting in front of each other, engaging in scripted conversation. This led to fewer movements, gestures and body shifts throughout the dataset.

The lack of available datasets for personality prediction from non-verbal interactive behaviour, complicates the necessary benchmarking evaluation. Two datasets that could potentially be used to evaluate this task are [42, 43], where the portrayed interactions happen in a more
natural way, people interacting with each other standing and not following a pre-scripted dialogue. However, the SALSA dataset [42] gives only self-assessment personality annotations, which is not suitable for the task of APP. Moreover, the recordings, involving 18 subjects during an indoor social event, show all subjects simultaneously from an overhead perspective, resulting in noisy and messy data. While the AICO corpus [43] presents a very similar dataset to the MHHRI (dyadic interaction and messy data. While the AICO corpus [43] presents a very similar dataset to the MHHRI (dyadic interaction recorded from a side-view RGB Kinect sensor of two people standing in front of each other, among others), it does not provide complete BIG5 personality annotations.

Comparing the performance of the four networks presented in our work with other approaches of DL for APP, reviewed in [11, 12], would be cumbersome as none of the previous work have been trained for the same task. However, [24] and [15] provide an evaluation of the MHHRI dataset. They evaluated the classification performance using an SVM with a function kernel in conjunction with First Person Vision and Second Person Vision individual features. A similar evaluation is performed in [14], on the ELEA corpus, using ridge regression and linear SVM regression classifiers. The performance of the better classification model, in these studies, varies for each trait. The classification methods, datasets and metrics differs significantly between studies, thus not allowing us to provide a complete and fair comparison. Nonetheless, conscientiousness was the most successfully predicted personality trait by our models (62% balanced accuracy for VGG DAN+ and 3DResNet). This value is higher than the accuracy values reached by [24,15] and, most significantly, by [14] where their results for openness to

Table 4. Comparison of the four architectures in terms of the precision (P), recall (R), $F_1$ score ($F_1$), balanced accuracy (Acc.) and ROC area under the curve (AUC) of the personality trait classification task.

| Metric | Architecture | Extroversion | Agreeableness | Conscientiousness | Neuroticism | Openness |
|--------|--------------|--------------|---------------|-------------------|-------------|---------|
| P      | 3DCNN        | 0.08         | 0.46          | 0.44              | 0.58        | 0.28    |
|        | 3DResNet     | 0.06         | 0.47          | 0.40              | 0.64        | 0.27    |
|        | VGG DAN+     | **0.80**     | 0.50          | **0.71**          | 0.59        | **0.28**|
|        | CNN+LSTM     | 0.07         | **0.55**      | 0.36              | 0.56        | 0.25    |
| R      | 3DCNN        | 0.05         | 0.63          | 0.35              | 0.75        | 0.24    |
|        | 3DResNet     | 0.02         | 0.50          | **0.51**          | **0.80**    | 0.19    |
|        | VGG DAN+     | **0.08**     | 0.61          | 0.36              | 0.75        | 0.14    |
|        | CNN+LSTM     | 0.07         | **0.66**      | 0.31              | 0.61        | **0.38**|
| $F_1$  | 3DCNN        | 0.06         | 0.53          | 0.39              | 0.65        | 0.26    |
|        | 3DResNet     | 0.03         | 0.49          | **0.55**          | **0.71**    | 0.22    |
|        | VGG DAN+     | **0.15**     | 0.55          | 0.48              | 0.66        | 0.19    |
|        | CNN+LSTM     | 0.07         | **0.60**      | 0.33              | 0.58        | **0.30**|
| Acc.   | 3DCNN        | 0.39         | 0.38          | 0.50              | 0.43        | 0.46    |
|        | 3DResNet     | 0.43         | 0.41          | 0.52              | 0.62        | 0.46    |
|        | VGG DAN+     | **0.54**     | 0.44          | **0.62**          | **0.54**    | **0.46**|
|        | CNN+LSTM     | 0.33         | **0.50**      | 0.44              | 0.41        | 0.40    |
| AUC    | 3DCNN        | 0.40         | 0.28          | 0.49              | 0.29        | 0.38    |
|        | 3DResNet     | 0.31         | 0.39          | **0.65**          | **0.49**    | **0.47**|
|        | VGG DAN+     | **0.45**     | 0.36          | 0.62              | 0.49        | 0.45    |
|        | CNN+LSTM     | 0.31         | **0.54**      | 0.44              | 0.40        | 0.36    |

Table 5. Comparison of the four architectures in terms of the macro, micro and samples average precision (P), recall (R), $F_1$ score ($F_1$), area under the ROC curve (ROC AUC), Hamming score, and subset accuracy.

| Architecture | Average | Micro avg. | Macro avg. | Samples avg. | Micro avg. | Macro avg. | Samples avg. | Micro avg. | Macro avg. | Samples avg. | Micro avg. | Macro avg. | Samples avg. | Micro avg. | Macro avg. | Samples avg. | Micro avg. | Macro avg. | Samples avg. | Micro avg. | Macro avg. | Samples avg. | Micro avg. | Macro avg. | Samples avg. |
|--------------|---------|------------|------------|--------------|------------|------------|--------------|------------|------------|--------------|------------|------------|--------------|------------|------------|--------------|------------|------------|--------------|------------|------------|--------------|------------|------------|--------------|
| 3DCNN        | P       | 0.44       | 0.47       | 0.45         | 0.48       | 0.49       | 0.13         | R          | 0.37       | 0.38       | 0.38       | 0.37       | 0.47        | 0.43         | $F_1$ | 0.50       | 0.49       | 0.48       | 0.48       | 0.47       | 0.44       | 0.49       | 0.13         |
| 3DResNet     | P       | 0.43       | 0.41       | 0.40         | 0.38       | 0.46       | 0.13         | R          | 0.51       | 0.47       | 0.47       | 0.47       | 0.52        | 0.43         | $F_1$ | 0.50       | 0.50       | 0.50       | 0.50       | 0.50        | 0.50       | 0.50       | 0.13         |
| VGG DAN+     | P       | **0.55**   | **0.45**   | **0.50**     | **0.57**   | **0.58**   | **0.23**    | R          | 0.41       | 0.41       | 0.40       | 0.40       | **0.53**    | **0.47**     | $F_1$ | 0.48       | 0.48       | 0.48       | 0.48       | 0.48        | 0.48       | 0.48       | 0.23         |
| CNN+LSTM     | P       | 0.58       | 0.39       | 0.40         | 0.40       | 0.40       | 0.23         | R          | 0.55       | 0.52       | 0.52       | 0.52       | 0.50        | 0.49         | $F_1$ | 0.49       | 0.49       | 0.49       | 0.49       | 0.49        | 0.49       | 0.49       | 0.23         |
experience, considered to be the most challenging trait to predict also by [15].

These results can be explained by taking into consideration two problems equally contributing to the performance of our architectures: the problem of unbalanced data and of the subjectiveness and lack of consistency in the annotations. As shown in Table 1, extroversion and openness to experience are the most unbalanced and less represented traits in the ‘high value’ class, while conscientiousness is one of the traits that is more equally balanced. This reinforces the theory that having unbalanced data resulted in under-performance. In addition, an analysis of the annotations provided in the original work describing the MHHRI dataset [24] showed that conscientiousness had the highest self-acquaintance agreement (similarity between the personality judgements made by self and acquaintances) among all traits. This means that the labels for this trait can be considered the most reliable in the dataset. Other traits, like openness to experience, presented low self-acquaintance agreement among the annotations, meaning that their labels cannot be considered as reliable as the ones of the conscientiousness trait. This underlines that the task of APP is challenging even among human annotators.

Overall, the VGG DAN+ and the 3DResNet outperform the other models, with an overall accuracy (average Hamming score) of 58% and 55%, respectively. Besides the overall accuracy scores, this finding is supported by them showing more consistency and relatively higher values across the personality traits, as seen in the ROC curves (Figure 6) and in the results in Tables 4 and 5. This is especially significant for the conscientiousness trait where they reached 62% in the balance accuracy. Finally, the hybrid CNN + LSTM model performs the worst. It performs inconsistently for all the trait predictions, achieving an average Hamming score of 45% and an average subset accuracy of 7%. Moreover, as underlined in Figure 6, it even underperforms a random classifier for most of the traits. We found that volumetric convolutions led to better performance for the APP task from the body language of human–human interactions than combining classical 2D convolutions and a LSTM network. In [32] an empirical study of the effects of different spatiotemporal convolutions for action recognition in video was presented and they found a noticeable gap between the performance of 2D and that of 3D or mixed convolutional models, suggesting that motion modelling is important for action recognition. Although interpretability for deep video architectures is still in its early stages and the DL community does not have a clear concept of how to decode spatiotemporal features, works such as [44] help us give insights as to why volumetric convolutions seem to work better for our task: on average, networks powered by 3D convolutions focus on shorter
and more specific sequences than networks using 2D convolutions and LSTM cells.

6. Conclusions and future work

In this paper, we analysed the effectiveness of body-related non-verbal cues and DL architectures for predicting apparent personality traits in social human–robot interactions. We customised 4 state-of-the-art DL architectures for the APP task on side-view videos from the MHHRII dataset.

If social robots could form a first impression of the personality of their users, based on the BIG5 personality traits, they could be enhanced with the ability of approaching and relating to their users in a more personalised way, adding value to the human–robot interaction itself, as the first encounter between a robot and a human can be crucial for both short-term engagement and long-term interactions [45].

Although the performance varied between models and personality traits, our evaluation showed the potential of the analysed architectures in predicting the different personality traits. These results are a starting point for discussing the benefits that using bodily signals features from video data input can have for APP, and its application to adaptive social robots.

In the future, we plan to use the AICO corpus [43], fully annotated with the BIG5 personality trait labels, to further verify these results, and to perform an empirical search for the best hyperparameters that will optimise the performance for these models. Moreover, our work sets a good starting point to study optimal strategies of adapting robots to the personality of human users in human–robot interactions. In psychology, various theories have argued on what could be the best personality match for each trait. Implementing our models in realistic human–robot setups will help to understand how these theories apply to robots and what are the traits their users expect to see, leading to an empirical analysis of building better human–robot companionship [46]. There have already been efforts in HRI to adapt the personality of robots to the personality of their users, demonstrating that this approach is beneficial for the overall interaction [47]. However, personality matching was achieved based on prior given measures of personality. This approach presents three problems: first, measures may not match with the actual personality of the person; second, they cannot be adapted; third, they cannot be used to interact with people robots are meeting for the first time. To that end, our work takes a step forward by providing an instrument for the community to automatically predict the apparent personality traits of robots’ users.

Note

1. https://gesture.chalearn.org/2016-looking-at-people-eccv-workshop-challenge.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was partially supported by a grant of AIST-AIRC (Japan) for the collaboration with the University of Manchester. The study is based on the results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO). The work was also supported by the EPSRC UKRI TAS Node on Trust and the European Research Council (H2020) projects PERSEO ETN and eLADDA ETN.

Notes on Contributors

Marta Romeo is a Postdoctoral Research Associate at the Cognitive Robotics Lab, Department of Computer Science, University of Manchester (UK). She is also a visiting researcher at the AI Research Center, AIST Tokyo Waterfront. She recently completed her Ph.D. under the supervision of Prof. Angelo Cangelosi on Deep Learning for Human–Robot Interaction in Elderly Care. Her research focuses on Human–Robot Interaction, with particular emphasis on evaluating the possibility of using deep learning methods to improve personalisation for Social Robots. Her current work is on Trustworthy Human–Robot Interaction and Autonomous Systems, within the UKRI TAS Node on Trust.

Daniel Hernández García is a Postdoctoral Research Associate at the Interaction Lab, School of Mathematical and Computer Sciences, Heriot-Watt University. Currently, he works in the SPRING EU project, focused on developing socially aware robots with the capacity for cognitive interactions in gerontological healthcare. His research has focused on Human–Robot Interaction and Social Robotics, Artificial Intelligence, Cognitive Systems, Interactive Task Learning, Machine Learning and Deep Learning, with the main interest in developing Intelligent and Socially-Aware Assistive Agents.

Ting Han is a research scientist at the Artificial Intelligence Research Center (AIST). Her research focuses on natural language and co-verbal gesture generation, aiming to enable robots and virtual agents to express themselves as humans do. She has worked on various research topics in this domain, including image caption generation, visually grounded natural language and hand-drawn sketch generation, as well as humorous text generation. She also serves as area chair and reviewer for top venues such as EMNLP, ACL, and INLG.

Angelo Cangelosi currently is Professor of Machine Learning and Robotics at the University of Manchester (UK). He also is Turing Fellow at the Alan Turing Institute London and Visiting Distinguished Fellow at AIST-AIRC Tokyo. His research interests are in developmental robotics, language grounding,
human robot-interaction and trust, and robot companions for health and social care. Cangelosi currently is Principal investigator for the ongoing projects ‘UKRI TAS Node on Trust’ (EPSRC, 2020–24, $1m), ‘THRIVE’ (US Air Force Lab, 2018–2023, $1m) and the H2020 projects TRAINCREASE, eLADDA ETN and PERSEO ETN. Overall, Cangelosi has secured over $32m of research grants as coordinator/PI, and produced more than 250 scientific publications. Cangelosi is Editor of the journals Interaction Studies and IET Cognitive Computation and Systems. His book ‘Developmental Robotics: From Babies to Robots’ (MIT Press) was published in January 2015, and recently translated in Chinese and Japanese. His latest book ‘Cognitive Robotics’ (MIT Press), coedited with Minoru Asada, will be published in 2021.

Kristiina Jokinen is a Senior Researcher at AI Research Center, AIST Tokyo Waterfront, and Adjunct Professor at the University of Helsinki. She is Member of the ELLIS Society and Life Member of Clare Hall at University of Cambridge. Her research concerns AI-based dialogue systems and multimodal dialogue modelling (speech, gaze and gestures) for human–robot interaction, and she has widely published on these topics, including four books. She has had a leading role in multiple national and international cooperation projects, and she led the development of WikiTalk, a Wikipedia-based robot dialogue system which won an award for Best Robot Design (Software Category) at the International Conference of Social Robotics 2017. She is currently leading research on trustworthy dialogue modelling in the EU-Japan collaboration project e-VITA. She is on the Advisory Board of H2020 project EMPATHIC and on the Steering Committee for the IWSDS dialogue workshop series. She has served as Chair or Senior Programme Member for authoritative conferences such as ICMI, ICSR, Interspeech, IVA, IWSDS, and SIGDial.

**ORCID**

Marta Romeo [http://orcid.org/0000-0003-4438-0255](http://orcid.org/0000-0003-4438-0255)

**References**

[1] Vinciarelli A, Mohammadi G. A survey of personality computing. IEEE Trans Affect Comput. 2014 Jul;5(3):273–291.

[2] Aly A, Tapus A. A model for synthesizing a combined verbal and nonverbal behavior based on personality traits in human–robot interaction. In: Proceedings of 8th ACM/IEEE International Conference on Human–Robot Interaction (HRI); 2013 Mar. IEEE press. p. 325–332. DOI:10.1109/HRI.2013.6483606.

[3] Klug M, Zell A. Emotion-based human–robot-interaction. In: 2013 IEEE 9th International Conference on Computational Cybernetics (ICCC); 2013. IEEE. p. 365–368. DOI:10.1109/ICCCyb.2013.6617620.

[4] Robert LP, Alahmad R, Esterwood C, et al. A review of personality in human–robot interactions. Found Trends Inf Syst. 2020;4(2):107–212. DOI:10.1561/2900000018.

[5] Breazeal C. Designing sociable robots. Cambridge, MA, USA: MIT Press; 2002.

[6] Subramanian R, Yan Y, Staiano J, et al. On the relationship between head pose, social attention and personality prediction for unstructured and dynamic group interactions. In: Proceedings of 15th ACM on International Conference on Multimodal Interaction (ICMI); 2013. Association for Computing Machinery, New York, NY, USA. p. 3–10. DOI:10.1145/2522848.2522862.

[7] Larsen RJ, Shackelford TK. Gaze avoidance: personality and social judgments of people who avoid direct face-to-face contact. Pers Individ Differ. 1996;21(6):907–917.

[8] Koppensteiner M. Motion cues that make an impression: predicting perceived personality by minimal motion information. J Exp Soc Psychol. 2013;49(6):1137–1143.

[9] Ng JYH, Hausknecht M, Vijayanarasimhan S. Beyond short snippets: deep networks for video classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); 2015. IEEE. p. 4694–4702. DOI:10.1109/CVPR.2015.7299101.

[10] Santhoshkumar R, Geetha MK. Deep learning approach for emotion recognition from human body movements with feedforward deep convolution neural networks. Procedia Comput Sci. 2019;152:158–165.

[11] Mehta Y, Majumder N, Gelbukh A, et al. Recent trends in deep learning based personality detection. Artif Intell Rev. 2019;53:2313–2339.

[12] Junior JCSJ, Güçlütürk Y, Perez M, et al. First impressions: A survey on vision-based apparent personality trait analysis. IEEE Trans Affect Comput. 2019. DOI:10.1109/TAFFC.2019.2874986.

[13] Okada S, Aran O, Gatica-Perez D. Personality trait classification via co-occurrence multiparty multimodal event discovery. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI); 2015. p. 15–22. DOI:10.1145/2818346.2820757.

[14] Ponce-López V, Chen B, Oliu M, et al. Chalearn LAP 2016: first round challenge on first impressions – dataset and results. In: Proceedings of the European Conference on Computer Vision (ECCV); 2016 Oct. Cham: Springer International Publishing. p. 400–418. DOI:10.1007/978-3-319-49409-8_32.

[15] Biel J, Gatica-Perez D. The YouTube lens: crowdsourced personality impressions and audiovisual analysis of Vlogs. IEEE Trans Multimedia. 2013;15(1):41–55.

[16] Zhang CL, Zhang H, Wei XS, et al. Deep bimodal regression for apparent personality analysis. In: Proceedings of the Computer Vision–ECCV 2016 Workshops; 2016. Cham: Springer International Publishing. p. 311–324. DOI:10.1007/978-3-319-49409-8_25.

[17] Subramanian A, Patel V, Mishra A, et al. Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features. In: Proceedings of the Computer Vision – ECCV 2016 workshops; 2016. Cham: Springer International Publishing. DOI:10.1007/978-3-319-49409-8_27.
multimodal apparent personality trait recognition. In: Proceedings of the Computer Vision – ECCV 2016 Workshops; 2016. Cham: Springer International Publishing. p. 349–358. DOI:10.1007/978-3-319-49409-8_28.

[21] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV); 2015. Santiago, Chile. IEEE. p. 4489–4497. DOI:10.1109/ICCV.2015.510.

[22] McCrea R. The five-factor model of personality. The Cambridge handbook of personality psychology; 2009. Cambridge, UK: Cambridge University Press. p. 148–161.

[23] Sanchez-Cortes D, Aran O, Mast M, et al. A nonverbal behavior approach to identify emergent leaders in small groups. IEEE Trans Multimedia. 2012 Jun;14:816–832.

[24] Çeliktutan O, Skordos E, Gunes H. Multimodal human–human–robot interactions (MHHRI) dataset for studying personality and engagement. IEEE Trans Affect Comput. 2017;10(4):481–497.

[25] Rammstedt B, John OP. Measuring personality in one minute or less: a 10-item short version of the big five inventory in English and German. J Res Pers. 2007;41(1):203–212.

[26] Willis J, Todorov A. First impressions: making up your mind after a 100-ms exposure to a face. Psychol Sci. 2006;17(7):592–598.

[27] He K, Gkioxari G, Dollár P, et al. Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV); 2017. IEEE. p. 2961–2969. DOI:10.1109/ICCV.2017.322.

[28] Waleed A. Mask R-CNN for object detection and instance segmentation on Keras and Tensorflow; 2017. Available from: https://github.com/matterport/Mask_RCNN.

[29] Lin TY, Maire M, Belongie S, et al. Microsoft COCO: common objects in context. In: Proceedings of the European Conference on Computer Vision (ECCV); 2014. Cham: Springer International Publishing. p. 740–755. DOI:10.1007/978-3-319-10602-1_48.

[30] He K, Zhang X, Ren S, et al. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV); 2015. Santiago, Chile. IEEE. p. 1026–1034. DOI:10.1109/ICCV.2015.123.

[31] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016. Las Vegas, USA. IEEE. p. 770–778. DOI:10.1109/CVPR.2016.90.

[32] Tran D, Wang H, Torresani L, et al. A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2018 Dec. Salt Lake City, USA. IEEE. p. 6450–6459. DOI:10.1109/CVPR.2018.00675.

[33] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of the International Conference on Learning Representations; 2015. Joshua Bengio and Yann LeCun, editors. San Diego, CA, USA.

[34] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–1780.

[35] Ullah A, Ahmad J, Muhammad K, et al. Action recognition in video sequences using deep bi-directional LSTM with CNN features. IEEE Access. 2018;6:1155–1166.

[36] Smith LN. A disciplined approach to neural network hyper-parameters: Part 1. Learning rate, batch size, momentum, and weight decay; 2018. Available from: arXiv:abs/1803.09820.

[37] Zhang M, Zhou Z. A review on multi-label learning algorithms. IEEE Trans Knowl Data Eng. 2014;26(8):1819–1837.

[38] Dembczyński K, Waegeman W, Cheng W, et al. Regret analysis for performance metrics in multi-label classification: the case of hamming and subset zero-one loss. In: Machine learning and knowledge discovery in databases; 2010. Berlin, Heidelberg: Springer. p. 280–295. DOI:10.1007/978-3-642-15880-3_24.

[39] Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manag. 2009;45(4):427–437.

[40] Brodersen KH, Ong CS, Stephan KE, et al. The balanced accuracy and its posterior distribution. In: Proceedings of the 20th International Conference on Pattern Recognition (ICPR); 2010. Istanbul, Turkey. IEEE. p. 3121–3124. DOI:10.1109/ICPR.2010.764.

[41] Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett. 2006;27(8):861–874.

[42] Alameda-Pineda X, Staiano J, Subramanian R, et al. Salsa: A novel dataset for multimodal group behavior analysis. IEEE Trans Pattern Anal Mach Intell. 2016;38(8):1707–1720.

[43] Ijuin K, Jokinen K. Exploring gaze behaviour and perceived personality traits. In: Social computing and social media. Design, ethics, user behavior, and social network analysis; 2020. Cham: Springer International Publishing. p. 504–512. DOI:10.1007/978-3-030-49570-1_35.

[44] Manttari J, Broome S, Folkesson J, et al. Interpreting video features: a comparison of 3d convolutional networks and convolutional LSTM networks. In: Proceedings of the Asian Conference on Computer Vision (ACCV); 2020 Nov. Cham: Springer International Publishing.

[45] Avelino J, Garcia-Marques L, Ventura R, et al. Break the ice: a survey on socially aware engagement for human–robot first encounters. Int J Soc Robot. 2021 Jan;1–27.

[46] Rossi S, Staffa M, Bove L, et al. User’s personality and activity influence on HRI comfortable distances. In: Proceedings of the International Conference of Social Robotics (ICSR). Springer International Publishing; 2017. Cham: Springer International Publishing. p. 167–177. DOI:10.1007/978-3-319-70022-9_17.

[47] Tapus A, Mătaric M. Socially assistive robots: the link between personality, empathy, physiological signals, and task performance. In: AAAI spring symposium; 2008 Jan. Stanford, CA, USA. AAAI. p. 133–140.