Toward an attentive robotic architecture: learning-based mutual gaze estimation in human-robot interaction

Maria Lombardi [1,*], Elisa Maiettini [1], Davide De Tommaso [2], Agnieszka Wykowska [2] and Lorenzo Natale[1]

[1] *Humanoid Sensing and Perception, Istituto Italiano di Tecnologia, via San Quirico 19D, 16163, Genoa, Italy*
[2] *Social Cognition in Human-Robot Interaction, Istituto Italiano di Tecnologia, via Enrico Melen 83, 16152, Genoa, Italy*

Correspondence*:
Maria Lombardi
maria.lombardi1@iit.it

ABSTRACT

Social robotics is an emerging field that is expected to grow rapidly in the near future. In fact, it is increasingly more frequent to have robots that operate in close proximity with humans or even collaborate with them in joint tasks. In this context, it is still an open problem the investigation of how to endow a humanoid robot with social behavioural skills typical of human-human interactions. Among the countless social cues needed to establish a natural social attunement, the paper reports our research towards the implementation of a mechanism for estimating gaze direction, focusing in particular on mutual gaze as fundamental social cue in face-to-face interactions. We propose a learning-based framework to automatically detect eye-contact events in online interactions with human partners. The proposed solution achieves high performance both *in silico* and in experimental scenarios. Our work is expected to be the first step towards an attentive architecture able to endorse scenarios in which the robots are perceived as social partners.

Keywords: mutual gaze, joint attention, human-robot interaction, humanoid robot, computer vision, experimental psychology, attentive architecture

1 INTRODUCTION

Joint attention (or shared attention) is one of the most important mechanisms occurring in a non-verbal interaction between two or more individuals. It is achieved when individuals direct their gaze on the same object or event in the environment as consequence of social gestures (e.g. gaze shift, pointing, facial expressions) (Moore et al., 2014). The ability to establish joint attention is crucial in many mechanisms of social cognition, for example comprehension, language development, intention, to cite a few (Tomasello, 1995; Tomasello et al., 2005; Mundy et al., 2007). A failure in such abilities, indeed, represents one of the earliest and basic social impairments in autism and communicative deficits (Mundy and Neal, 2000; Dawson et al., 2004).

In this context, designing and building an attention architecture enabling joint attention between a human and an embodied artificial agent, such as iCub, has inspired many researchers from different fields, spanning from artificial intelligence to robotics, from neuro and cognitive science to social science (Henschel et al., 2020; Wykowska, 2020). Inspired by the behaviour of human beings, our ambitious goal is to develop a robotic visual attention system that responds to several social cues characterising an effective non-verbal human interaction. For example, as social cue, eye gaze estimation plays a crucial role for the prediction of human attention and intention, and hence is indispensable for better understanding human activities (Kleinke, 1986; Emery, 2000). Humans, indeed, tend to look at an object before trying to grasp it with the hand (Voudouris et al., 2018). This implies that it is possible to predict human intention just observing where his/her attention is focused at.

In our long-range aim, the humanoid robot iCub will be able to establish social attunement with the human partner recognising and reproducing a wide range of social abilities in a human-like manner. The robot's ability to imitate human-like behaviours might bring the human to adopt the so called *intentional stance* as strategy towards the robot like s/he does with other humans (Marchesi et al., 2019). As proposed by the philosopher Daniel Dennett, intentional stance is the strategy of prediction and explanation that attributes beliefs, desires and intentions to an agent, and predicts its future behaviour from what it would be rational for an agent to do given those mental states (Dennett, 1989).

In this research report we present our first successful step in the ongoing implementation of such a robotic system. Specifically, we spent our initial effort on endowing iCub with the key ability of recognising eye-contact events. The report is organised in the following way. In the next section (Section 2) we discuss the importance of the mutual gaze in dyadic interactions. In Section 3 we describe the proposed solution for eye-contact detection. We benchmarked this algorithm in Section 4 where we compare it against the state-of-the-art. In Section 5, we test our architecture in a real HRI experimental setup, discussing the advantages of our solution in regard to the chosen case study. Finally, we draw the conclusion in Section 6.

2 FOCUS ON MUTUAL GAZE AND MOTIVATION

In the context of joint attention, eye-contact provides a foundation of effective social interaction since it signals the readiness for interaction and the attention of the partner. Given the sensitivity of a human when being watched by another one, it is not surprising that the mutual eye contact may influence the efficiency of the person-construal process (Macrae et al., 2002). For example, studies revealed that human observers are faster to detect target faces/eyes with direct gaze than those with averted gaze (Coelho et al., 2006) and the perceived eye-contact enhances the activation of components of the social brain network (Senju and Johnson, 2009).

While the effect of mutual eye gaze has been largely studied in human-human and human-screen scenarios with the use of reaction time measures (Galfano et al., 2012), saccadic behaviour (Ueda et al., 2014; Dalmaso et al., 2017a,b) and EEG (Hietanen et al., 2008; Pönkänen et al., 2011), few works exist in the literature investigating whether similar attention mechanisms arise in human-robot scenarios as well (Boucher et al., 2012).

For example in the context of human-human interaction, Chong et al. (2020) proposed a novel approach based on deep neural networks to detect eye contact in PoV camera video with reliability equivalent to expert human raters. The proposed algorithm has been used in this work as baseline for the comparison (see Section 4.3).

Wykowska (2021) underlined the importance of the role of humanoid robots as physical presence in real-time interaction since they provide higher ecological validity than screen-based stimuli and better experimental control than human-human interaction. Along the same line, Kompatsiari et al. (2018)

exploited the widely used Posner paradigm (Posner, 1980) to propose a novel interactive protocol involving the humanoid robot iCub (Metta et al., 2010) and examine the impact of mutual gaze on the mechanisms of joint attention.

Posner paradigm (together with its variations) is a neuropsychological test typically used to investigate attentional orienting in response to a directional cue. In such a gaze cueing task, the observer is typically asked to discriminate an object target (usually presented in a lateral location) while looking at a directional cue (e.g. schematic face or arrows) presented centrally, in-between the locations of potential target presentation. The cue can be either valid or invalid, depending on whether it pointed to the target object or to a different direction.

In their study, iCub was positioned between two lateral screens on which the object target was presented (in line with the Posner paradigm). iCub was used as the experimental apparatus both to establish real-time eye-contact with the human participant and to manipulate the directional gaze cue across the trials. The results revealed that the human reaction times depended on the combined effect of cue validity related to the iCub's gaze direction and social aspect of mutual gaze. Another example can be found in Stanton and Stevens (2017) where the Nao humanoid robot[1] was used to study the impact of three different levels of robot gaze (averted, constant and situational) in cooperative visual tracking task. Nevertheless the main drawback of the aforementioned studies is the use of the robot as a passive stimuli. Specifically, in both studies the humanoid robot was operated either with pre-programmed default text-to-speech and timed head movements or through pre-programmed gaze behaviour. As such, the robot had neither any perception of the real human's gaze nor any feedback by the surrounding environment.

Some authors support the notion that a robot embodying artificial models capable to reproduce human skills is a unique and invaluable tool to explain human cognition (Wykowska (2021); Pfeifer et al. (2007); Wainer et al. (2006)). With this motivation, in this work we propose a new module for iCub which allows to automatically detect whether the mutual gaze is established with the human partner during the interaction. Specifically, the report consists of three main contributions:

i. *Dataset collection for mutual gaze detection in frontal human-robot interaction*. In the context of frontal tasks, the dataset was collected general enough to be suitable in many different experimental scenarios. To the best of our knowledge it is the first mutual-gaze dataset collected involving a humanoid robot.

ii. *Designing, implementation and training of a learning module based on the aforementioned dataset*. Such a module is then embedded in the iCub's framework and validated both in silico and in online scenarios. Furthermore, we compare our method with the solution proposed in Chong et al. (2020) achieving an improvement in the accuracy of around 15 percentage points.

iii. As a case study, we select the experimental setup proposed in Kompatsiari et al. (2018) where iCub was used as a passive experimental apparatus. Within this framework, we performed several controlled experimental trials to test our application also in a time-constrained social robotics experiment.

Our approach aims at reducing the amount of hardware equipment required by the robot to detect mutual gaze with the human partner (e.g. external cameras, eye-tracker, and so on). The robot, indeed, relies only on the image frames captured by its eye-like cameras making the interaction as natural as possible. The algorithm developed in this work is an important building block for robotic setups that can be used to study human social cognition in naturalistic interaction.

[1] https://www.softbankrobotics.com/emea/en/nao

3 EYE-CONTACT LEARNING APPROACH

3.1 Data collection

3.1.1 Participants

A total of 24 participants were recruited for the data collection (mean age = 29.54 ± 3.14, 15 females). All participants had normal or corrected normal vision (6 participants out 24 wore glasses) and provided written informed consent. The data collection was conducted at the Istituto Italiano di Tecnologia, Genoa, and it was approved by the local ethical committee (Comitato Etico Regione Liguria).

3.1.2 Setup

The humanoid robot iCub embeds two Dragonfly2 cameras[2] (right and left eye); only one eye-camera was used with the frame resolution set to 640x480 pixels. In this study we used the right eye-camera, but the left-eye camera could be used equivalently. In order to have also higher quality images for the training phase of the proposed eye-contact classifier, a second dataset was also collected with the Intel RealSense depth camera D435[3] (see Figure 1 for a visual evidence). The RealSense camera was mounted on the iCub's head through a 3D printed headseat. The middleware YARP (Yet Another Robot Platform) (Metta et al., 2006) was used to integrate the different modules (e.g. iCub's controller, cameras, data dumper, code modules). The recording setup is shown in Figure 1. In line with what we claimed in Section 1 – i.e. to avoid need of external hardware – we underline that the RealSense camera was used only for acquiring training data. In the deployment phase, the system was always tested using images provided by the cameras mounted in the eyes of the iCub.

3.1.3 Task

Participants were asked to sit in front of the iCub at a distance of around one meter and to establish first mutual gaze and then averted gaze with the iCub's eyes in order to acquire frames both in eye-contact and in no eye-contact condition. In the eye-contact recording session, participants were also asked to look at the iCub's eyes but moving first their torso and then their head (Figure 1). For each position, the frame was captured both by the iCub's right camera and the RealSense pressing the bar space of the laptop's keyboard. The final datasets consist of 484 frames each (207 in eye-contact and 277 in no eye-contact condition).

3.2 Eye-contact classifier

Once the dataset was collected, the vector feature is extracted from each frame image by means of OpenPose[4] (Cao et al., 2019), a well-known real-time system for multi-human pose estimation. Specifically, OpenPose takes as an input a $w \times h$ color image as input and produces in output the 2D locations (x, y) of anatomical keypoints for each person in the scene with the corresponding detection confidence level k. Relying on a multi-stage deep convolutional neural network, OpenPose can jointly detect body, face, hands and foot keypoints reaching high accuracy and real-time performance, regardless the number of people in the image.

In our work, a subset of 19 face keypoints are considered (8 points for each eye, 2 points for the ears and 1 for the nose), resulting in a vector of 57 elements (i.e. the triplet (x, y, k) is taken for each point). Then, the

[2] http://wiki.icub.org/images/c/c9/POINTGREY_-_Dragonfly2.pdf

[3] https://www.intelrealsense.com/depth-camera-d435/

[4] https://github.com/CMU-Perceptual-Computing-Lab/openpose, https://github.com/robotology/human-sensing

detected keypoints are centered with respect to the head centroid, computed as the mean coordinates of all face keypoints, and normalised on the farthest point from the head centroid. The use of the face keypoints as feature vector has the main advantage of making the classifier independent of the light conditions and the picture's background.

The resulting feature vector is finally used as input to the binary classifier. Support Vector Machine (SVM) with RBF kernel was chosen to address this classification task. We compared the SVM with a random forest classifier; the former was chosen because it reported the best performance in terms of accuracy and F1-score (for a detailed comparison, see *Supplementary Material*). Moreover, given the results of the Principal Components Analysis (PCA), we considered the RBF kernel (see the *Supplementary Material* for further details). The hyperparameters of the SVM model were selected using an exhaustive search over a grid parameters and optimised by a 5-fold cross-validation (Pedregosa et al., 2011). After the training, the classifier's output is the pair (r, c) where $r = 1$ if mutual gaze is detected (0 otherwise), while $c \in [0, 1]$ is the confidence level of the prediction.

The overall learning architecture is depicted in Figure 2.

3.3 Training details

The mutual gaze classifier was trained both using the dataset collected with the RealSense and with iCub's eye. From now on, we refer to the classifier trained with the dataset from iCub's right eye since it reported higher performance metrics. For the full comparison between the two datasets, see *Supplementary Material*.

The acquired dataset was augmented in order to be robust to the degenerative case in which OpenPose fails to detect the eyes' boundaries and the pupils. To simulate such a condition, the coordinates of those keypoints in case of eye-contact were set to zero, while the others (namely, the ones for nose, ears and eyes) are left unchanged. Moreover, we applied a further augmentation by geometrically rotating the face keypoints, extracted by OpenPose, to the left and right of a certain angle around the face centroid to cover a wider range of head rotations (not covered by the acquired samples). In detail, facial keypoints were rotated to the left and right by an angle $\alpha \in \{15°, 30°, 45°, 60°\}$ taking the $\{5\%, 10\%, 10\%, 5\%\}$ of the data respectively. The final augmented dataset consist of 654 samples (377 in eye-contact, 277 in no eye-contact).

We handled the unbalanced dataset properly weighting each class of classification. Such weights were chosen inversely proportional to class frequencies in the input data.

Finally OpenPose parameters were tuned in order to have the best performance for the considered dataset (e.g. neural network resolution, images at different scales, and so on).

4 RESULTS

4.1 Evaluation on the collected test set

For the training of the classifier, the dataset was split into two subsets taking 19 out of 24 participants for the training set and the others 5 participants for the test set. The dataset was split $k = 5$ times in order to average the performance over different participants subsets and evaluate the statistical properties of the method. The performance were evaluated in terms of accuracy, precision, recall and F1-score reaching in all metrics values around 90%. Precisely we had: accuracy $= 0.91 \pm 0.03$, precision $= 0.90 \pm 0.08$, recall $= 0.89 \pm 0.06$, F1-score $= 0.89 \pm 0.04$.

4.2 Evaluation on temporal sequences

The mutual-gaze classifier was validated also on video streams recorded from the iCub's camera during different controlled interactions with a human. In detail, four video streams were recorded in order to cover the following scenarios: 1) no mutual gaze, 2) frontal mutual gaze, 3) human rotating the head to left/right while keeping mutual gaze with the robot, and 3) human rotating the torso while keeping mutual gaze with the robot. To avoid the flickering in the classifier predictions caused by the high video frame rates, we implemented a mechanism to propagate the predictions to those frames for which the classifier output is not available due to frame rate incompatibilities. The reason behind this is that, in practical settings, it is reasonable to assume coherent predictions in a $\sim 100ms$ time span. To this aim, we implemented a buffer of 3 elements at inference time. The actual classifier result was selected through a majority rule evaluated on the buffer. The implementation of the buffer allows to reach even higher level of accuracy. Specifically, the accuracy registered in the first three scenarios reaches its maximum value – i.e. 1.0 –, whereas in the last one the accuracy is 0.93. Analysing the last scenario, the classifier made wrong predictions when the human's torso reached the extreme angles of 90 (right) and -90 (left) while keeping the head straight toward the robot (see the videos in *Supplementary Material*). Such a drop in performance for the extreme torso rotations is reasonable, since the classifier was trained for frontal task.

4.3 Comparison with State-of-the-art method

In this Section the mutual gaze classifier is compared with the solution proposed in Chong et al. (2020). To the best of our knowledge, this is the most recent solution in the current literature that best adapts to our purposes. In Chong et al. (2020) authors trained a deep convolution neural network (i.e. ResNet-50 (He et al., 2016)) as backbone to automatically detect eye contact during face-to-face interactions. As network performance, authors reported an overall precision of 0.94 and F1-score of 0.94 on 18 validation subjects. The network was trained only with egocentric cropped frames of the individuals' face.

Because the training code of Chong et al. (2020) was not released by the authors, we used the publicly available pre-trained model. We tested this model on our scenario where the participants wore face-masks due to Covid19's ordinance and the frames captured by the robot were low quality frames. Since the algorithm used in Chong et al. (2020) failed to detect the bounding boxes of the humans' face in 33% of cases (probably due to the face-masks), we used OpenPose for the bounding box detection. Such bounding box was then used to crop the image sent as input to the convolution neural network. This was done to obtain a fair comparison between the two algorithms. Accuracy and F1-score were evaluated as metrics both on the test set and on the video streams:

- **Proposed approach**
  - *Test set*. Accuracy $= 0.91 \pm 0.03$; F1-score $= 0.89 \pm 0.04$.
  - *Stream videos*. Accuracy $= 0.97$; F1-score $= 0.98$.
- **Chong et al. (2020) + OpenPose**
  - *Test set*. Accuracy $= 0.76 \pm 0.05$; F1-score $= 0.77 \pm 0.06$.
  - *Stream videos*. Accuracy $= 0.89$; F1-score $= 0.82$.

Since data were normally distributed (Shapiro-Wilk test, p-value > 0.05), paired T-test was performed to assess the statistical difference between the performance of the two approaches (accuracy: p-value $= 0.01$, Cohen's d $= 2.009$, 95% CI for Cohen's d $[0.385, 3.581]$; F1-score: p-value $= 0.037$, Cohen's d $= 1.375$, 95% CI for Cohen's d $[0.072, 2.609]$).

On the test set we obtained an improvement of 15% in accuracy and of 12% in F1-score, whereas on the video streams we obtained an improvement of 8% in accuracy and of 6% in F1-score. In addition, our method is based on a low dimensional feature vector computed from facial and body landmarks. With respect to Chong et al. (2020), and other methods based on RGB information, it can be trained with less expensive hardware and without acquiring sensitive information (i.e. full RGB images depicting faces) from subjects.

The drop in the performance reported by Chong et al. (2020) in their work demonstrates the need of collecting a new dataset and shows that the current approaches in literature are not suitable for our scenario. Indeed, the considered setting is challenging both for the presence of face masks and for the low resolution camera that often is available in humanoid robots. On the contrary authors in Chong et al. (2020) used high resolution camera from camera glasses (1080p resolution). Notably, we could not compute the performance of our algorithm on the dataset used in Chong et al. (2020), because the latter was not made publicly available due to constraints imposed by the IRB protocol.

4.4 Model interpretability

With the aim of understanding which face keypoints have larger contribution to the final output of the learning architecture, SHAP analysis was performed on the trained SVM model. SHAP (SHapley Additive exPlainations) is a method based on coalitional game theory used to explain individually each prediction made by the learning algorithm. For each individual prediction, a value (SHAP value) is assigned to each feature as measure of its impact on the model's output. The final contribution for each feature is evaluated averaging its SHAP values over a set of predictions (Lundberg and Lee, 2017).

In Figure 3 the bar plot of the feature impact on the model output is reported for the first 20 most important face keypoints. It can be observed that the internal points of the eyes (pts $15, 16, 38, 39, 40, 42$) and partially the ears (pt 18) have a mean SHAP value between 0.02 and 0.09; this means that a change in these features in input has an impact on the prediction of around $2 - 9\%$ percentage points. The analysis reveals that there is no feature that predominates on the others but all the elements of the feature vector make a comparable contribution to the prediction in output. This is also confirmed by the principal components analysis reported in the *Supplementary Material*. The PCA performed on the data, indeed, does not make any improvement to the system implying that none of the considered features is completely redundant.

5 DEPLOYMENT IN AN EXPERIMENTAL SETUP

Next, we further validated our approach presented in the Section 3. As testbed example, we integrated our algorithm in the experimental scenario presented in Kompatsiari et al. (2018). In such a setup, participants were seated face-to-face with the iCub robot at a desk 125cm wide. iCub was positioned between two lateral screens on which target letters were presented to the participant. Also, iCub's height was set at 124cm from the floor in order to have its eyes aligned with participants' eyes (Figure 4).

The conclusions of Kompatsiari et al. (2018) were based on the assumption that mutual gaze was established between subjects and the robot, as confirmed by manual annotation by an experimenter. Therefore, the solution presented here offers a significant advancement, as it provides an automatic mechanism that can avoid manual annotation and implements a contingent robot behaviour allowing bi-directional eye contact mechanisms, which, as shown by the results of Kompatsiari et al. (2018), are crucial for establishing joint attention in HRI.

The experimental trial was designed as follows:

- iCub starts with the head pointing down and with its eyes closed for 2s;
- it opens its eyes for 500ms without moving the head;
- iCub looks towards the participant's eyes (eye contact) for 2.5s;
- iCub moves the head laterally towards one of the lateral screens, where the letter V or T appeared randomly either on the same screen where the robot is looking at (valid trial) or on the opposite screen (invalid trial) for 200ms;
- the participant is instructed to identify the target letter pressing V or T on the keyboard while keeping mutual gaze with the robot and without gazing at the screen.

To validate the classifier, we asked a total of 4 participants to carry out 8 blocks of 8 trials each. The experiments were controlled in order to have the ground truth for each block of trials. In detail, the participant was asked to *maintain* mutual gaze with the robot in 5 blocks of trials and to *always* simulate a distracted participant in the other 4 blocks left (e.g. checking the phone, looking at the lateral screens). To assure the quality of the ground truth, the experimenter monitored online eye movements of the participants and the trials were further checked offline before the analysis. Only one trial was discarded.

As done before, the performance were evaluated in terms of accuracy, precision, recall and F1 score. We registered: accuracy $= 0.97$, precision $= 0.95$, recall $= 1.00$, F1-score $= 0.97$.

6 CONCLUSION

In this research report we presented our first results of an ongoing work aiming at developing a novel attentive architecture for the humanoid robot iCub. In this context, we focused on the social cue of the mutual gaze making iCub capable of recognising eye-contact events while interacting online with a human partner. We validated the proposed mutual gaze classifier both computationally and experimentally, showing high performance values. We also compared the proposed approach with the state-of-the-art method Chong et al. (2020) reporting a consistent improvement in performance. We underline that our method requires neither any additional hardware (e.g. external camera, eye tracking glasses) nor a robot with embedded high-quality and expensive eye-cameras. Another advantage of our method is that it uses relatively low dimensional features extracted by facial landmarks which are intrinsically anonymous. With respect to other methods that use RGB information it can be re-trained with less expensive hardware and without storing personal data from subjects. Our results may potentially allow the research community to use an active robotic framework in more complex interactive scenarios helping the study of human cognition. For example, it has been previously found that the mutual gaze condition increases the level of engagement and/or rewarding during a human-robot interaction compared to averted gaze (Kampe et al., 2001). Similarly, Schilbach et al. (2010) investigated the neural correlates of joint attention finding that following or directing someone else's gaze activates several cortex areas of the brain related to the coordination of perceptual and cognitive processes.

Improving and extending the mutual gaze scenario to the wider problem of the gaze estimation is part of the current research. As a potential improvement, temporal information (e.g. temporal coherence between consecutive frames, optical flow, and so on) from dynamic data, like videos, could bring additional information to the system increasing performance and generalisation capabilities. Furthermore, the implementation of an attention system with the ability to detect social cues is a fundamental step toward the realisation of socially capable humanoid robots.

DATA AVAILABILITY STATEMENT

The anonymised data that support this study, the code and the learning trained models can be found at `https://github.com/hsp-iit/mutual-gaze-detection.git`. Further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Comitato Etico Regione Liguria. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

AW and LN conceived the main idea of the study. ML, EM and LN conceived and designed the learning architecture. ML collected the data, implemented the learning system, performed the experiments and analysed the data. ML, DDT deployed the algorithm on the iCub robot. ML, EM, DDT, AW and LN discussed the results. ML wrote the manuscript. All authors revised the manuscript.

FUNDING

This work has received support from the European Research Council under the European Union's Horizon 2020 research and innovation programme, ERC Starting grant ERC-2016-StG-715058, awarded to Agnieszka Wykowska. The content of this paper is the sole responsibility of the authors. The European Commission or its services cannot be held responsible for any use that may be made of the information it contains.

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ACKNOWLEDGMENTS

We thank Dr Giulia Pasquale for useful and insightful discussions and Serena Marchesi for her support and assistance in recruiting the participants for the data collection. We also thank all the people who took part in the experiments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: [Link Supplementary Material]

REFERENCES

Boucher, J.-D., Pattacini, U., Lelong, A., Bailly, G., Elisei, F., Fagel, S., et al. (2012). I reach faster when i see you look: gaze effects in human–human and human–robot face-to-face cooperation. *Frontiers in neurorobotics* 6, 3

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y. (2019). Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence* 43, 172–186

Chong, E., Clark-Whitney, E., Southerland, A., Stubbs, E., Miller, C., Ajodan, E. L., et al. (2020). Detection of eye contact with deep neural networks is as accurate as human experts. *Nature communications* 11, 1–10

Coelho, E., George, N., Conty, L., Hugueville, L., and Tijus, C. (2006). Searching for asymmetries in the detection of gaze contact versus averted gaze under different head views: a behavioural study. *Spatial vision* 19, 529–545

Dalmaso, M., Castelli, L., and Galfano, G. (2017a). Attention holding elicited by direct-gaze faces is reflected in saccadic peak velocity. *Experimental Brain Research* 235, 3319–3332

Dalmaso, M., Castelli, L., Scatturin, P., and Galfano, G. (2017b). Trajectories of social vision: Eye contact increases saccadic curvature. *Visual Cognition* 25, 358–365

Dawson, G., Toth, K., Abbott, R., Osterling, J., Munson, J., Estes, A., et al. (2004). Early social attention impairments in autism: social orienting, joint attention, and attention to distress. *Developmental psychology* 40, 271

Dennett, D. C. (1989). *The intentional stance* (MIT press)

Emery, N. J. (2000). The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & biobehavioral reviews* 24, 581–604

Galfano, G., Dalmaso, M., Marzoli, D., Pavan, G., Coricelli, C., and Castelli, L. (2012). Eye gaze cannot be ignored (but neither can arrows). *Quarterly Journal of Experimental Psychology* 65, 1895–1910

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*

Henschel, A., Hortensius, R., and Cross, E. S. (2020). Social cognition in the age of human–robot interaction. *Trends in Neurosciences* 43, 373–384

Hietanen, J. K., Leppänen, J. M., Peltola, M. J., Linna-Aho, K., and Ruuhiala, H. J. (2008). Seeing direct and averted gaze activates the approach–avoidance motivational brain systems. *Neuropsychologia* 46, 2423–2430

Kampe, K. K., Frith, C. D., Dolan, R. J., and Frith, U. (2001). Reward value of attractiveness and gaze. *Nature* 413, 589–589

Kleinke, C. L. (1986). Gaze and eye contact: a research review. *Psychological bulletin* 100, 78

Kompatsiari, K., Ciardo, F., Tikhanoff, V., Metta, G., and Wykowska, A. (2018). On the role of eye contact in gaze cueing. *Scientific reports* 8, 1–10

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, eds. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc.), vol. 30

Macrae, C. N., Hood, B. M., Milne, A. B., Rowe, A. C., and Mason, M. F. (2002). Are you looking at me? eye gaze and person perception. *Psychological science* 13, 460–464

Marchesi, S., Ghiglino, D., Ciardo, F., Perez-Osorio, J., Baykara, E., and Wykowska, A. (2019). Do we adopt the intentional stance toward humanoid robots? *Frontiers in psychology* 10, 450

Metta, G., Fitzpatrick, P., and Natale, L. (2006). Yarp: yet another robot platform. *International Journal of Advanced Robotic Systems* 3, 8

Metta, G., Natale, L., Nori, F., Sandini, G., Vernon, D., Fadiga, L., et al. (2010). The icub humanoid robot: an open-systems platform for research in cognitive development. *Neural networks : the official journal of the International Neural Network Society* 23, 1125–34. doi:10.1016/j.neunet.2010.08.010

Moore, C., Dunham, P. J., and Dunham, P. (2014). *Joint attention: Its origins and role in development* (Psychology Press)

Mundy, P., Block, J., Delgado, C., Pomares, Y., Van Hecke, A. V., and Parlade, M. V. (2007). Individual differences and the development of joint attention in infancy. *Child development* 78, 938–954

Mundy, P. and Neal, A. R. (2000). Neural plasticity, joint attention, and a transactional social-orienting model of autism. In *International review of research in mental retardation* (Elsevier), vol. 23. 139–168

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, 2825–2830

Pfeifer, R., Lungarella, M., and Iida, F. (2007). Self-organization, embodiment, and biologically inspired robotics. *science* 318, 1088–1093

Pönkänen, L. M., Alhoniemi, A., Leppänen, J. M., and Hietanen, J. K. (2011). Does it make a difference if i have an eye contact with you or with your picture? an erp study. *Social cognitive and affective neuroscience* 6, 486–494

Posner, M. I. (1980). Orienting of attention. *Quarterly journal of experimental psychology* 32, 3–25

Schilbach, L., Wilms, M., Eickhoff, S. B., Romanzetti, S., Tepest, R., Bente, G., et al. (2010). Minds made for sharing: initiating joint attention recruits reward-related neurocircuitry. *Journal of cognitive neuroscience* 22, 2702–2715

Senju, A. and Johnson, M. H. (2009). The eye contact effect: mechanisms and development. *Trends in cognitive sciences* 13, 127–134

Stanton, C. J. and Stevens, C. J. (2017). Don't stare at me: the impact of a humanoid robot's gaze upon trust during a cooperative human–robot visual task. *International Journal of Social Robotics* 9, 745–753

Tomasello, M. (1995). Joint attention as social cognition. *Joint attention: Its origins and role in development* 103130, 103–130

Tomasello, M., Carpenter, M., Call, J., Behne, T., and Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences* 28, 675–691

Ueda, H., Takahashi, K., and Watanabe, K. (2014). Effects of direct and averted gaze on the subsequent saccadic response. *Attention, Perception, & Psychophysics* 76, 1085–1092

Voudouris, D., Smeets, J. B., Fiehler, K., and Brenner, E. (2018). Gaze when reaching to grasp a glass. *Journal of vision* 18, 16–16

Wainer, J., Feil-Seifer, D. J., Shell, D. A., and Mataric, M. J. (2006). The role of physical embodiment in human-robot interaction. In *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication* (IEEE), 117–122

Wykowska, A. (2020). Social robots to test flexibility of human social cognition. *International Journal of Social Robotics* 12, 1203–1211

Wykowska, A. (2021). Robots as mirrors of the human mind. *Current Directions in Psychological Science* 30, 34–40

FIGURE CAPTIONS

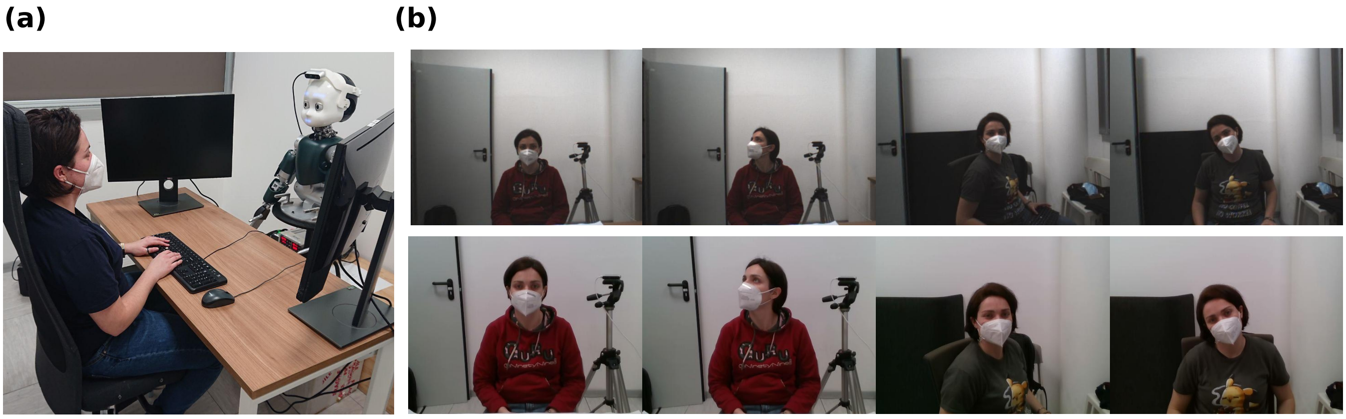

Figure 1. Dataset collection. (a) Overall setup. The participant was seated at a desk in front of iCub. The latter was mounted with a RealSense camera on its head. **(b)** Sample frames were recorded using both iCub's camera (first row) and the RealSense camera (second row). Different frames capture different human positions (rotation of the torso/head) and conditions (eye-contact and no eye-contact).

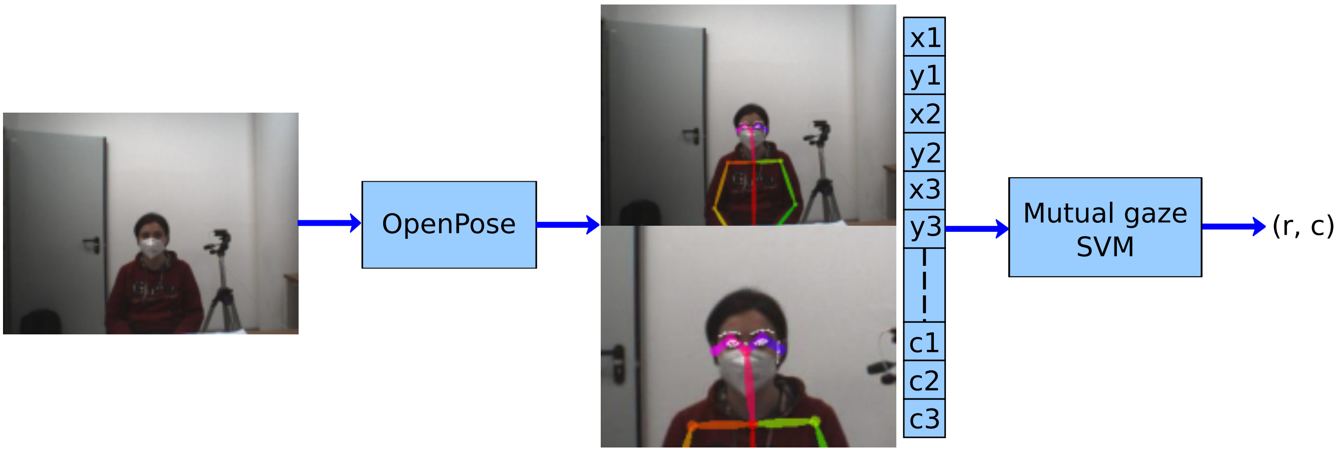

Figure 2. Learning architecture. The acquired image is first used as input for OpenPose in order to get the facial keypoints and build the feature vector for the individual in the scene. Then, such a feature vector goes in as input to the mutual gaze classifier whose output is the pair (r, c), where r is the binary result of the classification (eye-contact/no eye-contact) and c is the confidence level.

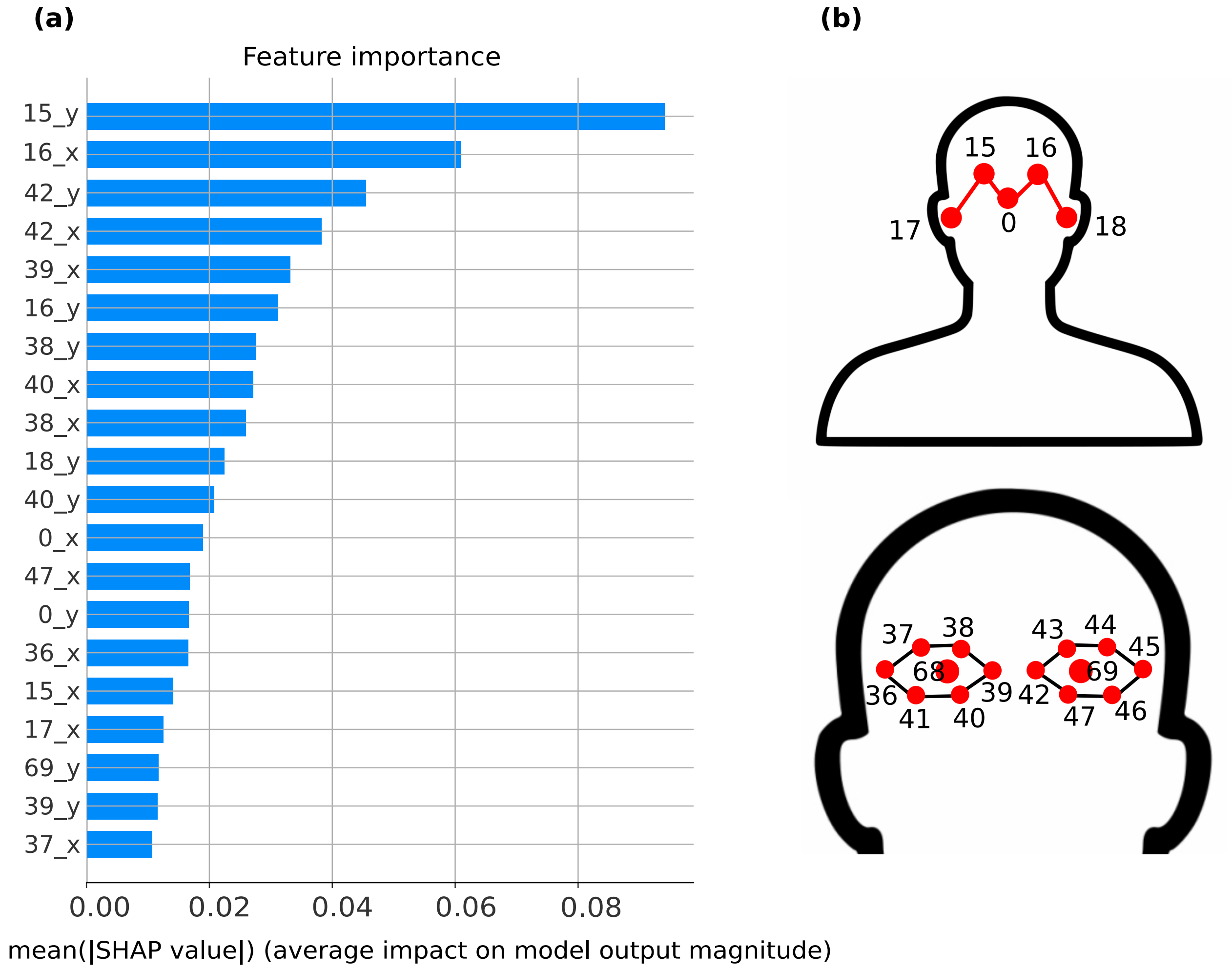

Figure 3. Feature importance. (a) Bar plot reporting on the x-axis the SHAP feature importance in percentage measured as the mean absolute Shapley value. Only the first 20 most important features are reported on the y-axis. **(b)** Numbered face keypoints of the feature vector.

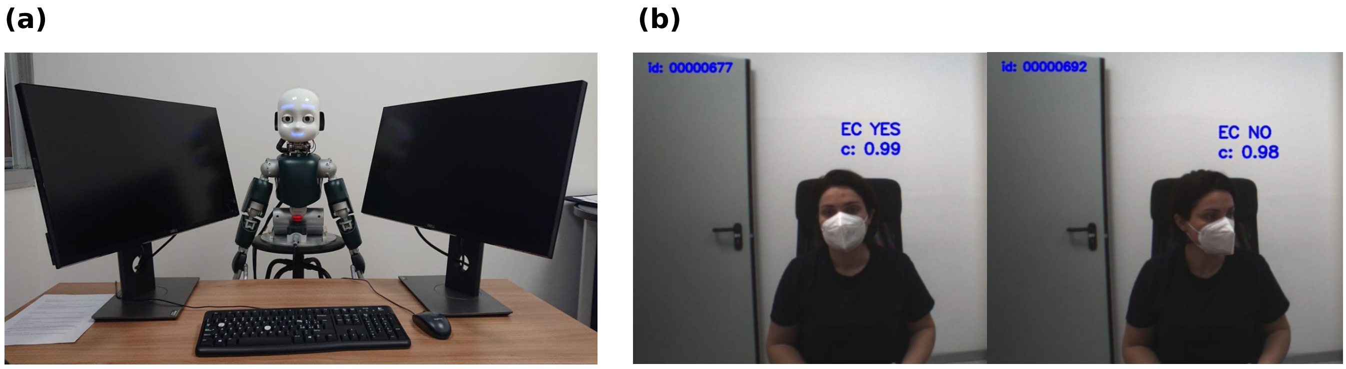

Figure 4. Experimental setup. (a) The iCub is positioned between two lateral screens face to face with the participant at the opposite sides of a desk that is 125cm wide. **(b)** Sample frames acquired during the experiment in which the participant first looks at the robot to make an eye contact and then simulates a distraction looking at the lateral screen. On each frame, the prediction (eye Contact yes/no) with the confidence value c is also reported.