The Gaze Dialogue Model: Nonverbal Communication in HHI and HRI

Mirko Raković, *Member, IEEE*, Nuno Ferreira Duarte, *Graduate Student Member, IEEE*, Jorge Marques, Aude Billard, *Fellow, IEEE*, and José Santos-Victor, *Fellow, IEEE*

***Abstract*—When humans interact with each other, eye gaze movements have to support motor control as well as communication. On the one hand, we need to fixate the task goal to retrieve visual information required for safe and precise action-execution. On the other hand, gaze movements fulfil the purpose of communication, both for reading the intention of our interaction partners, as well as to signal our action intentions to others. We study this *Gaze Dialogue* between two participants working on a collaborative task involving two types of actions: 1) *individual action* and 2) *action-in-interaction*. We recorded the eye-gaze data of both participants during the interaction sessions in order to build a computational model, the *Gaze Dialogue*, encoding the interplay of the eye movements during the dyadic interaction. The model also captures the correlation between the different gaze fixation points and the nature of the action. This knowledge is used to infer the type of action performed by an individual. We validated the model against the recorded eye-gaze behavior of one subject, taking the eye-gaze behavior of the other subject as the input. Finally, we used the model to design a humanoid robot controller that provides interpersonal gaze coordination in human–robot interaction scenarios. During the interaction, the robot is able to: 1) adequately infer the human action from gaze cues; 2) adjust its gaze fixation according to the human eye-gaze behavior; and 3) signal nonverbal cues that correlate with the robot's own action intentions.**

***Index Terms*—Bio-inspired robotics, human-in-the-loop, human-robot interaction (HRI).**

Manuscript received 23 February 2022; revised 27 June 2022 and 29 September 2022; accepted 1 November 2022. This work was partially supported by the RBCog-Lab PINFRA/22084/2016, EU MSCA Project ACTICIPATE, FCT Funding ISR-Lisboa/LARSyS LA/P/0083/2020, by the Lisbon Ellis Unit (LUMLIS), and the Nuno Ferreira Duarte Ph.D. Grant with Reference PD/BD/135116/2017. This article was recommended by Associate Editor H. A. Abbass. *(Mirko Raković and Nuno Ferreira Duarte contributed equally to this work.) (Corresponding author: Mirko Raković.)*

Mirko Raković is with the Vislab, Institute for Systems and Robotics, Instituto Superior Técnico, Universidade de Lisboa, 1049-001 Lisbon, Portugal, and also with the Faculty of Technical Sciences, University of Novi Sad, 21000 Novi Sad, Serbia (e-mail: rakovicm@isr.tecnico.ulisboa.pt).

Nuno Ferreira Duarte is with the Vislab, Institute for Systems and Robotics, Instituto Superior Técnico, Universidade de Lisboa, 1049-001 Lisbon, Portugal, and also with the LASA, Swiss Federal Institute of Technology Lausanne, 1015 Lausanne, Switzerland (e-mail: nferreiraduarte@isr.tecnico.ulisboa.pt).

Jorge Marques and José Santos-Victor are with the Vislab, Institute for Systems and Robotics, Instituto Superior Técnico, Universidade de Lisboa, 1049-001 Lisbon, Portugal (e-mail: jsm@isr.tecnico.ulisboa.pt; jasv@isr.tecnico.ulisboa.pt).

Aude Billard is with the LASA, Swiss Federal Institute of Technology Lausanne, 1015 Lausanne, Switzerland (e-mail: aude.billard@epfl.ch).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TCYB.2022.3222077.

Digital Object Identifier 10.1109/TCYB.2022.3222077

I. Introduction

HUMANS can routinely engage in joint actions, and coordinate their movements with others in very sophisticated manners. Such interactions occur in situations as diverse as cooking, cleaning, assembling complex structures, carrying heavy loads, or performing team sports. There is a continuous adaptation between the interaction partners to each other's actions, in closed loop. These tasks involve a collaborative process to coordinate attention, communication, and actions to achieve a common goal [1]. During this process, humans observe the behavior of their partners to anticipate their actions, and to plan their own actions accordingly. As the humans' neural mechanisms of motor preparation are relatively slow, prediction allows to significantly accelerate the dynamics of our reactions [2]. This is fundamental to enhance coordination in humans and in human–robot interactions (HRIs).

The perception of eye gaze movements is particularly important for action coordination [3] and to anticipate the intentions of others [4]. When working on a collaborative task, the human eye gaze alternates between looking at each other's eyes, seeking the confirmation and engagement of the counterpart, and fixating the goal position before and during reaching actions [5]. Sebanz and Knoblich [6] reported that the ability to gaze at the right location in a timely manner substantially enhances coordination with other individuals. In infant–parent relationship, the eye-gaze communication works as a tool to study the infant's development [7], [8].

In this article, we study what we call the *Gaze Dialogue* between two people working on a joint task (Fig. 1), involving a sequence of actions where each person is either a leader or a follower. After one action is completed, the roles are changed, that is, the leader becomes a follower, and vice-versa. The process is repeated until the entire task is finished. We consider two types of actions: 1) *individual action*, *placing* an object, and 2) *action-in-interaction* [9], *giving* an object to someone. After the experiments, we analyze the eye-gaze of each pair of participants (Section III and the *Gaze Dialogue Model* learns the interdependency/coordination between the leader's and follower's eye-gaze movements (Section IV).

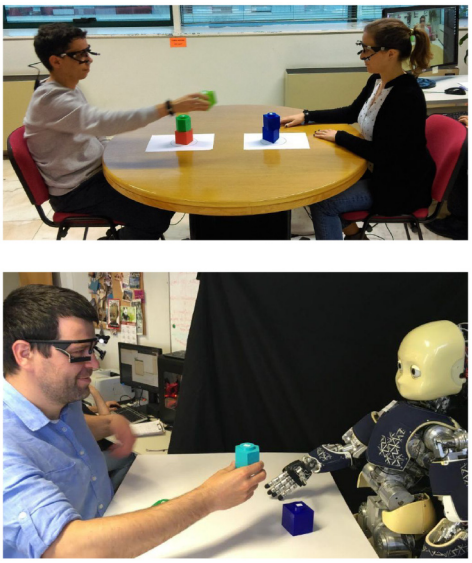

Fig. 1. Two humans are performing a task of assembling two towers, without any verbal communication. The experiment requires them to execute individual actions, pick and place that is, *placing*, or handover, that is, *giving*. On the bottom image, a human is performing the same task as before, but interacting with a robot with human-like behavior.

The *Gaze Dialogue Model* combines two key functionalities: 1) predicting the gaze fixations of others and planning one's own fixations and 2) using the gaze fixations to predict the actions of others and to plan/generate one's own actions. The performance of the *Gaze Dialogue Model* is initially assessed by comparing the model results against the data acquired in human–human interaction (HHI) experiments. We use the fixations and actions performed by one of the subjects in the HHI as the model input, in order to predict the fixations and actions of the other human in the interaction. In the next step of performance validation, we implement the model in a humanoid robot controller, which drives the robot eye fixations, during the HRI experiments (Section V). The robot controller, based on the *Gaze Dialogue Model*, takes the human gaze fixations as the input to predict the human next gaze fixations and actions, while at the same time, generating its own appropriate gaze fixations and planning its own actions. The results show the robot successfully identifying the actions of the human partner, and acting in a manner that is consistent with the HHI scenario. The behavior of the robot is described quantitatively and it can be visualized in the supplementary material in Section V-C. Finally, in Section VI, we draw some conclusions and establish directions for future work.

II. State of the Art

Gaze behavior has drawn substantial research interest for over 30 years in order to fathom the intricacies of interaction of living things [10]. In the last decade, the research efforts have been directed to HRI and how it can improve robot's inclusion in modern societies.

Yäcel et al. [11] described a joint attention model between a robot and a human, by analyzing the human's head pose. The authors mentioned that head orientation was easier to track and to estimate, and that the analysis of the experimenter's eyes would require stable and high-resolution image acquisition, with a prohibitive computational cost. To detect human deceptiveness, Yu et al. [12] analyzed the head orientation, facial expressions, to study how people tend to synchronize nonverbal cues during the interaction. Instead, we use the eye-gaze motion instead of head movements alone, because eye-gaze has been shown to generate more accurate results in terms of gaze synchronization during interaction [13], [14]. Ivaldi et al. [15] developed a controller for the humanoid robot iCub that "reads" the head orientation to understand which object is being fixated by a human. This was achieved with an external RGB-depth camera, but the controller was limited to detecting colored objects only. The main drawback of these approaches is that head orientation is taken as an approximation to the actual eye-gaze fixations. For a better interaction experience, the robot should detect the interaction partner, and use eye contact to coordinate its actions with the human counterpart. Chadalavada et al. [16] utilized the head-mounted eye-tracker data to analyze trajectories and eye-gaze patterns in a scenario where humans have to navigate around robots. The robot projects the directional movement intention and from eye-gaze data it is possible to quantify which projection approach is most effective. Notwithstanding, none of these approaches use eye-gaze data for action recognition and generation. Instead, we apply, in real-time, the human eye-tracker information to interact with a robot in an HRI setup.

Some approaches were capable of using the full potential of the human eye analysis in HRI scenarios. Pereira et al. [17] studied gaze to measure if the robot is socially present, that is, if the robot is looking at the task or the person. Humans classifying a robot as socially present improves overall likeness and perceived intelligence. Similarly, Kompatsiari et al. [18] evaluated the impact of humanoid robots eye-contact with humans in an HRI and showed that it also improves sociality and quality of interaction. However, other works [19] concluded that fixating an object of shared attention rather than gazing at the robotic partner, is the most meaningful prediction of engagement. Palinko et al. [20] worked on the detection of facial features, such as the pupil position in the eye, that were used to estimate the direction of the human eye gaze. Their approach used the robot RGB cameras, and did not require any extra sensors to detect the human gaze direction. Unfortunately, the cameras suffered from poor resolution, sensitivity to varying light conditions, and was limited to distinguish left and right orientation of human eye-gaze fixations. Domhof et al. [21] integrated eye-gaze estimates, extracted with eye tracking systems, in a robot controller to understand the fixation point of humans. The fixation point was calculated either from the eye tracker or from pointing gestures to relevant objects. This information was provided to a robot controller and, using the robot's RGB-depth camera, it was possible to distinguish the object that the human was aiming at. Currently, the research using eye-trackers are limited to understanding if a person is fixating at an object or not [22], [23]. In [24], we implemented human-inspired eye-gaze cues on robots to achieve human-level legibility of robot actions. In this work, we use eye-tracking data to study the joint attention between two humans, by analyzing the contextual gaze information extracted from a dataset combining two eye trackers for an HHI scenario [25].

Andrist et al. [26] focused on bidirectional gaze interaction between a human and a virtual agent in a sandwich-making task. This interaction was modeled with an HHI experiment and it demonstrated how gaze nonverbal cues can lead to a faster completion time of the task and reduce the error rate. Notwithstanding, one of the limitations was the fact that the experiments were only applied to the "instructor role," that in this article we designate as the "leader" perspective. In our work, we study not only the behavior of the participant who is performing the actions but also of the participant who is observing or interacting.

There are recent works using gaze shift behavior that studied the impact of the leading participants' attention [27], as well as head mounted eye-tracking which studied the human gaze focus in virtual environments [28]. The focus of our experiments will involve the latter techniques for detecting human behavior in social interactions. Lukic et al. [29] studied the individual actions performed by humans during pick-and-place operations. The motion of the hand, arm, head, and eyes was tracked when a person moved an object from point A to point B. The data was used to model the inner coupling between the different human body part movements. The individual actions of Lukic et al. [29] are similar to the *placing* actions in this article. The gaze behavior provided a reference signal to the visual servoing module, by shifting from one goal position (initial location of the object), to the final position (final destination of the object). Schydlo et al. [14] developed a learning-based action anticipation model based on motion and gaze fixation data. It showed that it can improve the prediction time of actions by introducing human gaze behavior. In this work, we want to predict the action while it is being performed in real time. In addition, instead of relying on raw pixel information, by contrast, our approach identifies the contextual information, that is, the object of interest we are looking at, the face of the other subject, etc. In addition, in this work, we model a *Gaze Dialogue* which predicts the actions from the perspective of both parties involved in the interaction.

Previous datasets such as the ones used in [14], [24], and [30] only provide the gaze behavior from the perspective of the subject performing the actions (actor). Instead, the dataset [25] used in this work encompasses human gaze behavior in a dyadic interaction scenario from two perspectives: 1) the perspective of the subject leading the action (*leader*) and 2) the subject who is observing/interacting with the first subject (*follower*). The dataset in [31] provides human eye-tracker and skeleton motion data in a scenario performing complex tasks such as cleaning a table with multiple objects. Ondras et al. [30] gathered a dataset of audio samples from human speech and motion from the head, hand, and torso. It looked into the nonverbal communication of humans during speech and applied it to a humanoid robot for more expressive and relatable robots. The network would output the appropriate joint angles for the robot to perform similar nonverbal cues. The authors argue that robots should exhibit human-like motion behavior while having dialogues with humans. Our dataset overcomes several limisierrantations of past datasets by including full-body motion and gaze data in manipulation tasks set in 3-D world spaces. However, our work extends [24] by introducing a second agent as well as interactive actions such as handovers between subjects. Hence, we believe that, although we do not provide full-body motion, the dataset includes the most meaningful dimensions in human behavior: eye-gaze, head orientation, and arm motion of two people at the same time.

The contribution of this work is to build a model of the eye-gaze communication system that can: 1) predict the gaze behavior and actions of the person we are interacting with and 2) generate our eye-gaze fixations as well as plan the action we will perform. For this purpose, the next section proceeds to analyze the eye-gaze measurements from two humans while they interact with each other.

III. Human–Human Interaction Experiment

The objective of this experiment is to study the human eye-gaze in its functional and communicational role in the context of interaction and action execution. In order to acquire the data necessary to build the *Gaze Dialogue* model, we designed an HHI experiment and recorded the eye-gaze data of two actors working on a collaborative task.[1] The experimental setup and the data acquisition procedure are detailed in the dataset paper [25], and only a brief description is provided here.

A. HHI Experiment Description

The experiment consists of a dyadic interaction task for assembling two towers from a stack of three objects placed next to each participant (top image of Fig. 1). The description of the task and the stack of objects are occluded from the other participant at the beginning of each round. In order to complete the task, the actors are required to perform a series of simple actions. For every action, each actor is either a *leader* or a *follower*, and these roles alternate for the subsequent actions. When the task starts, the leader picks the first object from the stack. If the object is a match for his/her tower, the leader will place it, while the follower only observes the action. In the case the object is for the other tower, the leader gives the object to the follower who will place it on his/her tower.

There are two types of actions the actors execute: 1) *placing* an object or 2) handing over, that is, *giving*, an object. In social psychology, these two actions elicit two different behaviors: 1) offline social cognition where the follower observes, and the leader receives no information from the follower and 2) online social cognition where only the follower adapts to the leader [32]. These behaviors are addressed in this article in the context of a human–robot scenario.

Once the towers are assembled, the task is updated, and a new round with new instructions for each participant starts. There is a total of four rounds, that is, each actor in a dyad assembles four different towers. The towers configuration and associated actions alternate each round, in order to prevent the participants from anticipating movements based on the previous experience.

[1]The dataset can be found in the following public repository: http://vislab.isr.tecnico.ulisboa.pt/datasets/#acticipate. More details about the dataset (e.g., data collection and motion capturing) are available upon contact via nferreiraduarte@isr.tecnico.ulisboa.pt.

B. Data Acquisition Setup and Dataset

To capture their eye movements, both participants wore a Pupil-Labs binocular gaze tracker [33]. The Pupil Labs binocular gaze tracking glasses are equipped with three cameras. Two infrared cameras record at 120 Hz and the third RGB camera provides a video stream of the egocentric view at 60 Hz. The movement of participants' head and arm were recorded using the Optitrack motion tracking system, running at 120 Hz. To record head gaze, we attached five reflective markers on each eye-tracking glasses. Each group of five markers define one rigid body transformation, whose position and orientation in the reference frame are recorded.

The lab streaming layer (LSL) [34] library is used to synchronously capture the data from sensors (motion capture and two eye-trackers) of each dyad. For LSL, we developed a Motive2LSL application that captures the broadcasted location of the markers and rigid bodies in the Motive software as well as the application that receives the data measurements from the two Pupil-Labs glasses.

C. Labeling and Analysis of Acquired Data

The described HHI experiment was repeated with three dyads, that is, six participants[2] (five males and one female, 22–45 years old, a mixture of academic undergraduates, graduates and staff employees, most unfamiliar with robots). All participants were naive to the objective of the experiment. Assembling one tower requires three actions, and each participant built four towers. Thus, the recordings include three actions for four towers for six participants, as a leader (i.e., 72 actions), and an equal number of actions as a follower. The recordings are paired, that is, each action is observed from both the leader's and the follower's perspectives.

The gaze fixations, illustrated in Fig. 10, are estimated and provided by the official software of the Pupil Labs gaze tracker. Each gaze fixation detected for each action counts as one frame of the video stream of the egocentric view (60 Hz). After data collection, it is visually inspected by an expert, in order to define the regions-of-interest, that is, the most frequent gaze fixations in the vicinity of meaningful "things" in the experiment. For a leader, the following fixations are considered: brick (B), follower's face (FF), follower's hand (FH), leader's own hand (LOH), follower's tower (FT), and leader's own tower (LOT). The fixations defined for a follower are: leader's face (LF), leader's hand (LH), leader's tower (LT), and follower's own tower (FOT). Gaze fixations which landed outside the previously mentioned regions-of-interest count, representing less than 1% of the total dataset, were considered as outliers and discarded in the analysis. Finally, these regions-of-interest are used to manually label both the leader's and follower's eye-gaze fixations. Table I shows one example of the leader's gaze fixation labeling process for a *giving* and a *placing* action.

[2]The ACTICIPATE studies do not include frail subjects and the subjects are recruited within the academic population at IST. All subjects are informed, adult, healthy young individuals, and there is absolutely no invasiveness in the tests.

TABLE I
EXAMPLES OF LEADER'S GAZE BEHAVIOR FOR EACH ACTION WITH TOTAL DURATION IN VIDEO FRAMES FOR EACH REGION OF INTEREST

Giving	Labels	B	FH	FT	FH	LOH	FT	FF	FT
	Duration (frames)	143	7	23	7	21	6	38	29
Placing	Labels	B	LOT	FF					
	Duration (frames)	78	31	8					

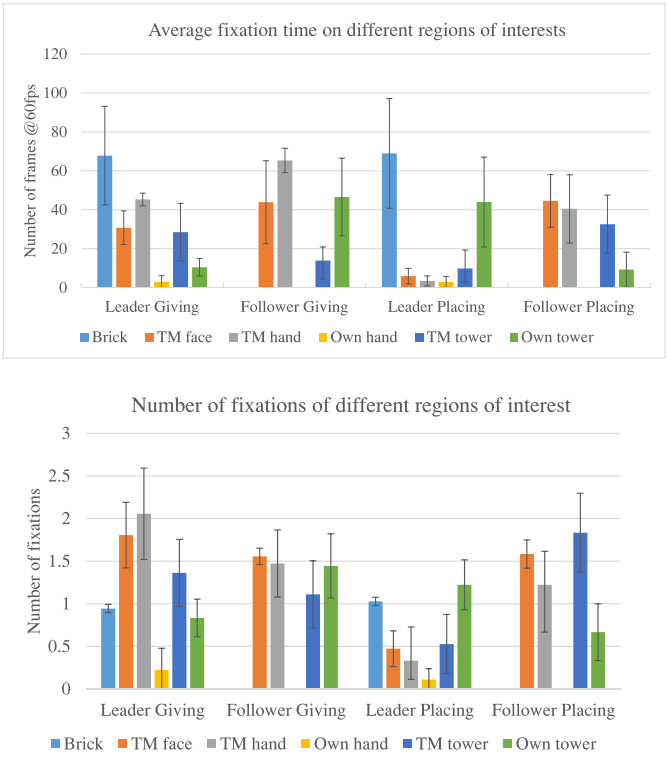

Fig. 2. Analysis of gaze fixations during HHI experiment. (a) Duration of gaze fixation for different regions of interest per action. (b) Number of gaze fixations of different regions of interest per action. TM stands for Team-mate.

Besides the gaze fixations, the significant events of an action are also annotated: "action start," "object picked," "object handed over" (only exists for the giving action), "object placed," and "end of action." Fig. 2 shows the average duration of gaze fixations toward different regions of interest across 72 actions for both types of actions, and in the plot below, it shows an average number of gaze fixations for identified regions of interest.

From Fig. 2, we can conclude that for the *giving* action, the leader has multiple gaze fixations, and the gaze fixation time is longer compared to the *placing* action. In the *placing* action, instead, the leader focuses mainly on his/her tower, whereas for the *giving* action, the leader switches several times between FF, hand and tower, and fixates those regions of interest for a significant amount of time. The leader fixates the brick evenly in the two actions.

The follower's regions of interest are different from the leader, that is, the brick does not exist, and looking at his/her hand after the brick is handed over was negligible. The follower's gaze fixation behavior is comparable between the *giving* and *placing* actions while there is a significant difference for the leader's gaze fixations. This is due to the follower's attempt to understand the leader's action. As a

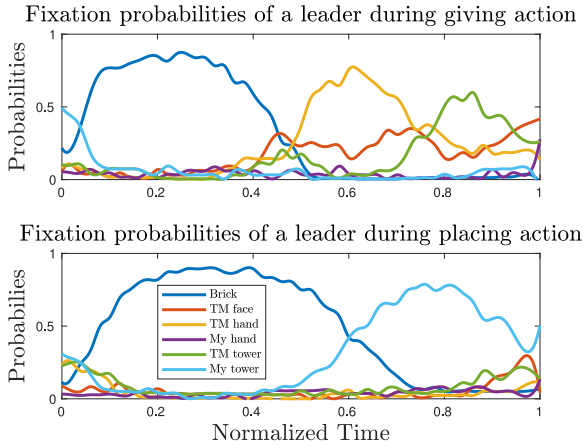

Fig. 3. Fixation probabilities of leader's gaze for giving (top figure) and placing (bottom figure) action.

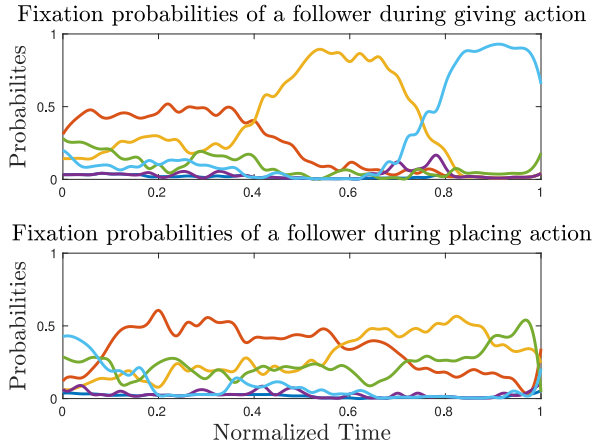

Fig. 4. Fixation probabilities of follower's gaze for giving (top figure) and placing (bottom figure) action.

result, we see the follower spend a significant amount of time fixating the LF and/or hand presumably attempting to "read" the action. The main difference is the number and duration of gaze fixations between the leader's and his/her tower. This occurs when the follower is aware of the action, fixating the goal, that is, either his/her tower, for visually controlling the *giving* action, or the LT, monitoring the execution of the *placing* action.

Fig. 3 shows the computed probabilities of the leader's gaze fixations over time, for both *giving* and *placing* actions, averaged for all actions in the dataset. The (empirical) probability was estimated by calculating the relative frequency of each gaze fixation over time, after normalizing the time-duration of all actions. In *giving*, the leader starts by fixating the brick then successively fixates the FH, the FT and, finally, the face. In *placing*, the leader fixates the brick first, and then his/her own tower, almost until the end of the action. Fig. 4 shows the follower's gaze fixations when observing the leader performing either a placing or a giving actions. The most notable fixations are TM Hand, and My Tower, which are predominant during a giving action. These occur when the goal of the follower is to grasp the object from the LH and to place it on his/her tower. At the beginning, until about 50% of the total time, the most probable fixation is TM Face, which indicates that the follower is trying to decode the leader's action intention. Given that in a placing action, the follower is not active, there is not one, but several probable fixations, reflecting a more passive role. Although at the end of the action, the follower fixates the LT (TM Tower) when it becomes clear that it is a placing action.

The main conclusions from this analysis are four-fold.

1) It is possible to predict the leader's action from the gaze fixations.
2) There is a clear distinction between the leader and the follower's gaze fixations (Figs. 3 and 4).
3) From the leader's perspective, there is a considerable focus on the brick, which is negligible in the follower's case. The difference may be justified by the roles of each subject in the experiments, the leader needs to manipulate the brick to complete the action, whereas the follower only needs to follow the leader's behavior.
4) The follower's gaze fixations were similar for both actions. This was not the case with the leader. The explanation may be related to the distinct nature of each action: an action-in-interaction (handover) requires communicating the intent; instead, an individual-action (placing a brick on the tower) does not.

Concerning the follower, the nature of the action is initially unknown, the behavior is similar for both actions until the moment when the action intention is understood. Once the follower decodes the action, the gaze behavior changes accordingly, which may justify the slight change in the tower fixations for the two actions. In the next section, we will present the model that learns from the gaze fixations to predict the leader's actions.

IV. Gaze Dialogue Model

The *Gaze Dialogue Model* integrates the eye-gaze communication that occurs during an interaction between two humans, with the arm-motor actions which result from the interaction. We start with a general model that represents each human as a separate system, with eye-gaze and arm-motor movements, together with the interpersonal links of nonverbal communication. The eye-gaze is used for predicting the fixations of others while, at the same time, generating one's fixations. Understanding the gaze fixations, we can infer the associated actions. The arm-motor cues represent the action of each actor. It predicts the actions of others, while, at the same time, it plans one's actions, generating the appropriate motor commands to complete the action.

The proposed *Gaze Dialogue Model* block diagram is shown in Fig. 5. The states are defined as the gaze fixation S_k and type of action A_k, for each actor, at time instant k. The model is composed of the Gaze Fixation system, identified by the blue blocks, and the Action Anticipation system, the yellow blocks. The Gaze Fixation system is responsible for predicting the fixations of others, and generating one's own fixations. The role of the Action Anticipation system is to predict the actions of others, and to plan one's own action.

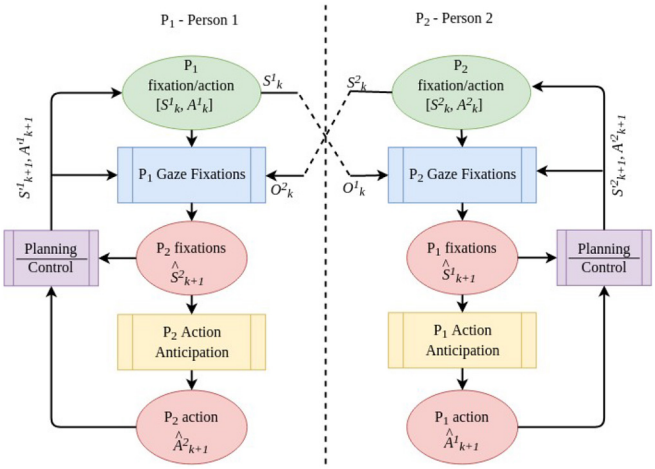

Fig. 5. Block diagram of the proposed general *Gaze Dialogue* model.

The *Gaze Dialogue Model* uses the history of a person's gaze fixations and actions, together with the observations O_k of the gaze fixations of the other person. The Gaze Fixation system predicts the gaze fixation of the other person at time $k+1$, while the Action Anticipation system predicts the type of action performed by the second person. The predictions of the fixations and actions, together with one's fixations and actions, are eventually fed back to the Planning/Control system, identified by the purple block. This block is responsible for the person's next gaze fixation and which action to perform.

The choice of hidden Markov models (HMMs) as a modeling tool is primarily due to its low complexity and modest data requirements compared to other highly complex, data hungry models, as deep neural networks. Acquiring synchronized human eye-gaze data is costly and HMMs proved that we can use small amounts of data to predict the eye-gaze movements and actions of others and generate one's own eye-gaze movements and actions. Finally, the HMM allows us to naturally incorporate the interdependencies between concerning human motor control units, such as eye-gaze, head and arm movements, and can run in real-time for HRI.

The general approach of the *Gaze Dialogue Model*, for the Gaze Fixation System, is described in Section IV-A and Section IV-B describes the Action Anticipation system.

A. Gaze Fixations

We have modeled the *Gaze Dialogue* with an HMM, where each actor has an associated internal state variable: $S_k \in \{U_1, \ldots, U_N\}$ where $U_1, \ldots, U_N$ are the admissible state values, that is, fixations, and $k \in \{1, \ldots, T\}$ denotes the discrete-time instants. The actor has access to an instantaneous observation: $O_k \in \{V_1, \ldots, V_M\}$ where $V_1, \ldots, V_M$ are the fixations of the other actor. The two sequences (state and observation)

$$S = (S_1, \ldots, S_T), O = (O_1, \ldots, O_T)$$

are represented by the HMM $\lambda = (\pi, C, D)$ where π denotes the probability distribution of the state variable at time $k = 1$, $C = (c_{i,j})$ denotes the transition matrix, and $D = (d_{i,j})$ denotes the emission matrix [35]. Since we consider two actors, denoted by P_1 and P_2, the above sequences are duplicated

$$S_k^1 \in \{U_1^1, \ldots, U_N^1\} \; O_k^1 \in \{V_1^1, \ldots, V_M^1\}$$
$$S_k^2 \in \{U_1^2, \ldots, U_N^2\} \; O_k^2 \in \{V_1^2, \ldots, V_M^2\}$$

and two different HMMs are used to generate the state and observation sequences for each actor: $\lambda^1 = (\pi^1, C^1, D^1)$ and $\lambda^2 = (\pi^2, C^2, D^2)$. The joint probabilities of the state and observation sequences for the two actors are

$$P(S^1, O^1) = \prod_{k=1}^{T} c^1_{S^1_{k-1}, S^1_k} \cdot d^1_{S^1_k}(O^1_k)$$
$$P(S^2, O^2) = \prod_{k=1}^{T} c^2_{S^2_{k-1}, S^2_k} \cdot d^2_{S^2_k}(O^2_k).$$

In the perspective of actor P_1, we predict the fixation at time $k+1$ of P_2, $\hat{S}^2_{k+1}$, and generate his/her own next fixation, S'^1_{k+1}. In the perspective of P_2, it predicts the fixation of P_1, $\hat{S}^1_{k+1}$, and generates the next fixation, S'^2_{k+1}, of P_1.

B. Action Anticipation System

The Action Anticipation System has the twin-goal of predicting the actions of others and planning one's own actions, based on the gaze fixations of the actors. In order for actor $j \in \{1, 2\}$ to predict the action $\hat{A}^i_{k+1}$ of another actor, $i \neq j$, we combine the information related to the observed gaze fixations $\hat{S}^i_{k+1}$. Based on the current gaze fixation of the other actor, we use the action probabilities from Table IV, to update an exponential moving average

$$P^i_a(k+1) = (1-\alpha)P^i_a(k) + \alpha\delta(k) \quad (1)$$

where k refers to time, and α is a constant smoothing factor. $\delta(k)$ is the probability $P^i_a(k)$ of action a occurring when actor i is looking at gaze fixation S^i_k at time k. P^i_a is the probability of actor i performing action a. During the interaction, the *Gaze Dialogue Model* predicts the actions of others, and allows one to plan our own actions. The predicted action $\hat{A}^i_{k+1}$ of actor i is updated for every new gaze fixations $\hat{S}^i_{k+1}$ the actor i is gazing for each time $k+1$. At the same time $k+1$, the Action Anticipation System allows the planning of the action A'^j_{k+1} associated to actor j. The exponential moving-average mechanism ensures a smooth evolution of the action probabilities, and filters out spurious noisy measurements.

C. Gaze Fixations for the Leader–Follower Interaction

We then adapt the *Gaze Dialogue Model* to the leader-follower relation that is extracted from the experimental data. The HHI *Gaze Dialogue Model* contains both the eye-gaze communication and the arm movements, as in the general model, with the difference that the leader's action is predefined (instruction), and the purpose of the follower is to understand the action, *giving* or *placing*, and act accordingly. Fig. 6 shows the block diagram of the model with a few modifications to reflect the leader–follower experiments from our scenario. From Gallotti et al. [32] in a leader–follower scenario, the leader leads the action, while the follower adapts its

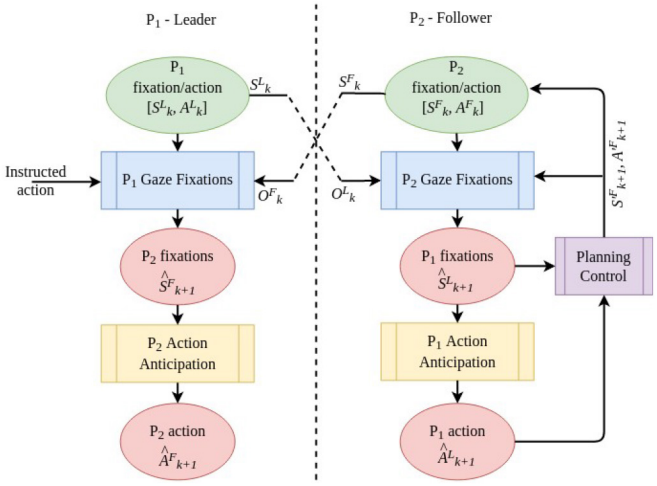

Fig. 6. Block diagram of the leader–follower *Gaze Dialogue* model.

behavior to match the leader's intention. As such, in the *Gaze Dialogue* model, the leader's block system is not in closed-loop, since the leader is not influenced by the follower. Since the action is instructed to the leader, the leader has to generate his/her own gaze fixations and action, without having to predict the interaction partner's gaze fixations or actions. On the other hand, the follower "reads" the leader's nonverbal cues (arm movements and gaze fixations) to infer the leader's action and, consequently, prepare his/her own action and provide the appropriate nonverbal cues.

The state and observation values match the labeled and leader/follower pair gaze fixations described in Section III. The leader has six different states, and the follower has four. For the leader/follower pair, we denote the states of, respectively, by S^L and S^F. The states and observations of the leader–follower

$$S^L_k \in \{U^L_1, \ldots, U^L_N\} \quad O^L_k \in \{V^L_1, \ldots, V^L_M\}$$
$$S^F_k \in \{U^F_1, \ldots, U^F_N\} \quad O^F_k \in \{V^F_1, \ldots, V^F_M\}$$

the HMMs for the leader–follower relation has the following parameters, $\lambda^L = (\pi^L, C^L, D^L)$, and $\lambda^F = (\pi^F, C^F, D^F)$.

For the leader–follower relation of the HHI experiments, the state and observation sequences are related, as the state of the leader becomes the observation of the follower, and vice-versa: $S^F = O^L$ and $S^L = O^F$. We can therefore define two sequences $S = (S_1, \ldots, S_T)$ and $O = (O_1, \ldots, O_T)$ and the leader has state sequence S and observation sequence O, while the follower has the opposite. The probabilistic models are

$$P(S, O) = \prod_{k=1}^{T} c^L_{S_{k-1}, S_k} \cdot d^L_{S_k}(O_k)$$
$$P(O, S) = \prod_{k=1}^{T} c^F_{O_{k-1}, O_k} \cdot d^F_{O_k}(S_k)$$

with different C and D matrices. The two HMMs are learned from the *giving* and *placing* actions data, and the obtained transition and emission matrices are given in Table II. In the Leader's perspective, the transition matrices size are 6×6, for the six states (B, FF, FH, LOH, FT, and LOT mentioned in Section III-C), and the emission matrices are 6×4, from the six leader's states to the follower's states (LF, LH, LT, and FOT). To note that FH and LH is the same as TM Hand in the perspective of the follower and leader, respectively. In the follower's perspective, there are only four states (LF, LH, LT, and LH); hence, the sizes are 4×4 and 4×6 for the transition and emission matrices, respectively.

There are four HMMs in total in the *Gaze Dialogue Model*, one for each person (leader versus follower) and for each action (placing versus giving). The HMMs are used by the leader to predict the follower's next state $\hat{S}^F_{k+1}$ and, conversely, by the follower to predict the leader's next state, $\hat{S}^L_{k+1}$. More important, by using posterior decoding, the follower can plan its next fixation S'^F_{k+1} in response to the leader's behavior. To measure the effectiveness of the gaze fixations' generated by the *Gaze Dialogue* model, we compare the generated follower's fixations against the real follower's fixations. These experiments involve, for each HHI dataset leader's input, to generate the follower's gaze fixations' output, and then compare to the real corresponding HHI follower's response. The accuracy is 67% ± 24% for the giving action, and 64% ± 24% for the placing action (1/6 for chance level). We can conclude from the accuracy results, which reflect the average and standard deviation for the HHI dataset, that the model is capable of generating a common follower's reaction taking into account the nondeterministic behavior during a leader–follower interaction. It is important to mention that the *Gaze Dialogue* model is not trying to mimic an exact leader–follower match for every possible variation.

Fig. 7 shows the leader–follower gaze fixations and the generated follower's fixations for placing and giving action sequences. The model takes the leader's gaze fixations (blue line on Fig. 7 top plot) as the input and estimates the predicted behavior of the follower using the posterior state probabilities (shown by the middle plot). The predicted gaze fixations of the follower (red line in Fig. 7 bottom plot) are the gaze fixations with the highest probability at each time instate. The follower's predicted gaze fixations are compared against the instance of the actual (recorded) gaze fixations (blue line in Fig. 7 bottom plot).

When analyzing the percentages of each gaze fixation for every action, in Table III, shows that the most common gaze fixations for each action are consistent with the behavior presented in Fig. 4. Although the recorded follower's gaze fixations for a single instance/specific action may differ from the predicted (probabilistic) gaze fixations, nonetheless, most of the time, the predictions match the observed fixations. The gaze fixations generated for the follower when performing a giving action are exactly the ones identified as the most fixated in the HHI dataset: 1) fixating the LF; 2) fixating the LH (both for decoding the human intention and for following the human arm trajectory); and 3) his/her own tower for placing the object and conclude the action. As for the placing action, the face of the leader is the most dominant, as shown in Fig. 4, followed by the LT which is when the leader is finalizing the placing action.

TABLE II
HMM PARAMETERS FOR THE LEADER (L) AND FOLLOWER (F) DEFINED BY TRANSITION MATRIX C AND EMISSION MATRIX D FOR *(G)iving* AND *(P)lacing* ACTIONS

	Leader	Follower
Giving	$C_G^L=\begin{bmatrix}0.9861 & 0.0016 & 0.0045 & 0.0016 & 0.0041 & 0.0020\\ 0.0018 & 0.9567 & 0.0316 & 0.0009 & 0.0081 & 0.0009\\ 0.0012 & 0.0241 & 0.9630 & 0.0006 & 0.0105 & 0.0006\\ 0 & 0 & 0.0541 & 0.9279 & 0.0180 & 0\\ 0.0019 & 0.0183 & 0.0144 & 0.0010 & 0.9587 & 0.0058\\ 0.0477 & 0.0053 & 0.0053 & 0.0186 & 0 & 0.9231\end{bmatrix}$ $D_G^L=\begin{bmatrix}0.4702 & 0.2838 & 0.1310 & 0.1151\\ 0.0799 & 0.5245 & 0.0477 & 0.3479\\ 0.1381 & 0.5705 & 0.0317 & 0.2597\\ 0.3091 & 0.4273 & 0.0818 & 0.1818\\ 0.0650 & 0.2367 & 0.0260 & 0.6723\\ 0.3837 & 0.2820 & 0.2238 & 0.1105\end{bmatrix}$	$C_G^F=\begin{bmatrix}0.9683 & 0.0209 & 0.0070 & 0.0038\\ 0.0064 & 0.9796 & 0.0030 & 0.0111\\ 0.0437 & 0.0198 & 0.9226 & 0.0139\\ 0.0030 & 0.0012 & 0.0012 & 0.9947\end{bmatrix}$ $D_G^F=\begin{bmatrix}0.6997 & 0.0396 & 0.1227 & 0.0217 & 0.0319 & 0.0843\\ 0.3022 & 0.1861 & 0.3626 & 0.0215 & 0.0832 & 0.0444\\ 0.6199 & 0.0752 & 0.0894 & 0.0183 & 0.0407 & 0.1565\\ 0.1818 & 0.1832 & 0.2449 & 0.0136 & 0.3507 & 0.0258\end{bmatrix}$
Placing	$C_P^L=\begin{bmatrix}0.9867 & 0.0008 & 0.0004 & 0.0008 & 0.0004 & 0.0109\\ 0.0098 & 0.9755 & 0.0098 & 0 & 0 & 0.0049\\ 0.0569 & 0.0081 & 0.9268 & 0 & 0.0081 & 0\\ 0 & 0 & 0 & 0.9722 & 0 & 0.0278\\ 0.0200 & 0.0114 & 0 & 0 & 0.9629 & 0.0057\\ 0.0064 & 0.0051 & 0.0013 & 0.0006 & 0.0064 & 0.9803\end{bmatrix}$ $D_P^L=\begin{bmatrix}0.4543 & 0.2213 & 0.2284 & 0.0960\\ 0.1387 & 0.4380 & 0.4161 & 0.0073\\ 0.2857 & 0.0408 & 0.1735 & 0.5000\\ 0.3299 & 0.4639 & 0.1856 & 0.0206\\ 0.1869 & 0.6920 & 0.0588 & 0.0623\\ 0.2286 & 0.4163 & 0.3367 & 0.0184\end{bmatrix}$	$C_P^F=\begin{bmatrix}0.9682 & 0.0143 & 0.0162 & 0.0012\\ 0.0090 & 0.9793 & 0.0104 & 0.0014\\ 0.0259 & 0.0121 & 0.9586 & 0.0035\\ 0.0180 & 0.0060 & 0.0419 & 0.9341\end{bmatrix}$ $D_P^F=\begin{bmatrix}0.7546 & 0.0131 & 0.0192 & 0.0220 & 0.0371 & 0.1540\\ 0.4273 & 0.0479 & 0.0032 & 0.0359 & 0.1597 & 0.3259\\ 0.5570 & 0.0575 & 0.0172 & 0.0182 & 0.0172 & 0.3330\\ 0.7250 & 0.0031 & 0.1531 & 0.0063 & 0.0563 & 0.0563\end{bmatrix}$

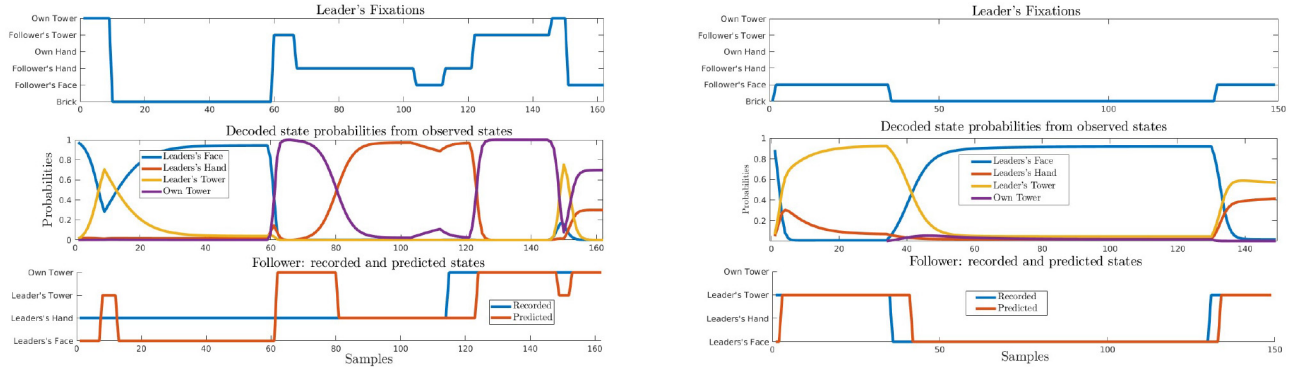

Fig. 7. Simulations of leader's and follower's internal model in the case when the leader's behavior during a giving action (left) and placing action (right). The top plot shows the leader's recorded gaze fixations, the middle plot the follower's fixation probabilities, and the bottom shows the follower's recorded and most likely fixations.

TABLE III
Gaze Dialogue FOLLOWER'S GAZE FIXATIONS PROBABILITIES

	LF	LH	LT	FOT
Giving	**0.42**	**0.32**	0.05	**0.21**
Placing	**0.61**	**0.13**	**0.20**	0.06

TABLE IV
PROBABILITIES FOR GIVING AND PLACING ACTION WITH RESPECT TO THE LEADER'S GAZE FIXATION

	Giving	Placing
Brick	0.496	0.504
Follower's face	0.841	0.159
Follower's hand	0.931	0.069
Own hand	0.520	0.480
Follower's tower	0.748	0.252
Own tower	0.186	0.814

D. Action Anticipation for the Human–Human Interaction

In the HHI experiments, only two actions are possible for the leader, *giving* and *placing*, or the follower, *receiving* and *not-receiving*. Taking into account the leader–follower relation, a *receiving action* is associated with a *giving* action and *not-receiving* to a *placing* action.

The prediction of a certain action combines the information related to the follower's current fixations, with the past probability of the same action. The action probabilities are

$$P_G(k+1) = (1-\alpha)P_G(k) + \alpha\delta(k) \quad (2a)$$

$$P_P(k+1) = (1-\alpha)P_P(k) + \alpha\delta(k) \quad (2b)$$

where P_G and P_P denote the probabilities of *giving* and *placing* action, respectively, for each time step k, and $\alpha = 0.05$. The update $\delta(k)$ depends on the values of Table IV, evaluated for each gaze fixation of the leader at time k. In the HHI experiment, the leader is "instructed" which action to perform (*giving* or *placing*). The action is unknown to the follower who needs to understand it from the nonverbal communication cues. The *Gaze Dialogue Model* infers the leader's action from the leader's eye-gaze fixations and, in turn, generate the follower's eye-gaze fixations and action.

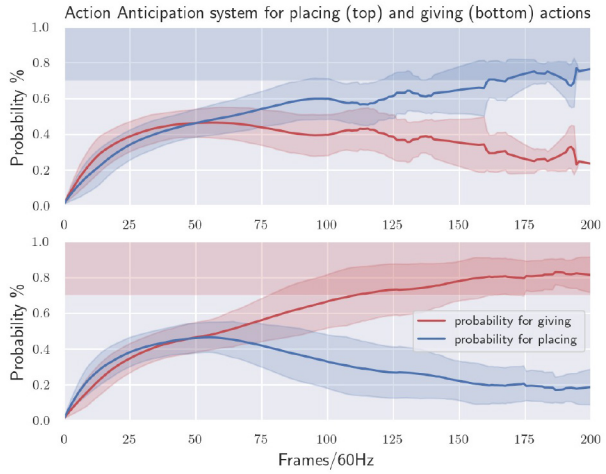

Fig. 8. Action anticipation results on the entire HHI dataset on classifying the actions as a placing or giving action, respectively.

The follower's Action Anticipation system uses the leader's observed gaze fixations. Each gaze fixation is associated with the probability to choose between two actions as given in Table IV. The probabilities were derived from the duration of each gaze fixation for each action, as given in Table I, divided by the total duration of gaze fixation throughout the HHI experiments. Table IV shows that the leader fixations at the brick or at his/her own hand are negligible for both actions, as the probabilities are close to 50%. Instead, other gaze fixations provide stronger gaze cues toward one of the two actions. The leader's gaze fixation at the FF, hand, or tower clearly communicates the intention of *giving* the brick, whereas gaze fixations at his own tower, presumably to visually guide the arm to properly place the brick, become strong cues for the *placing* action.

The Action Anticipation System is composed of two signals that represent the probabilities for the *giving* ($P_G(k)$) and *placing* ($P_P(k)$) actions, over time, with the initial values set to 50%. These signals are updated in each iteration. First, the action is selected based on the leader's current gaze fixation and the probabilities shown in Table IV. For example, if the leader's fixates the FF there is a 84.1% chance to select a *giving* action and a 15.9% to select a *placing* action. Based on the selected action, the δ values of (2a) and (2b) are updated to calculate the value of the signals $P_G(k)$) and $P_P(k)$ for the next time k. In case the leader gaze fixates the FF, and the *placing* action is selected, the δ of 0.159 is used for the signal $P_P(k)$ and -0.159 for the signal $P_G(k)$. The output signals $P_G(k)$ and $P_P(k)$ are smoothed with a moving average, and normalized with respect to the number of samples (i.e., the number of gaze fixations observed) collected up to time k. This approach is similar to a Markov Reward Process [36] that adds a reward signal to each state. In our case, the purpose is to decide which type of action, in order to prevent oscillatory behavior of the action prediction.

The accuracy of the Action Anticipation system are shown in Fig. 8. The action is classified either as *placing* or *giving* when the prediction reaches the region marked with a shaded color (set empirically as above 70%). A *giving* action can be

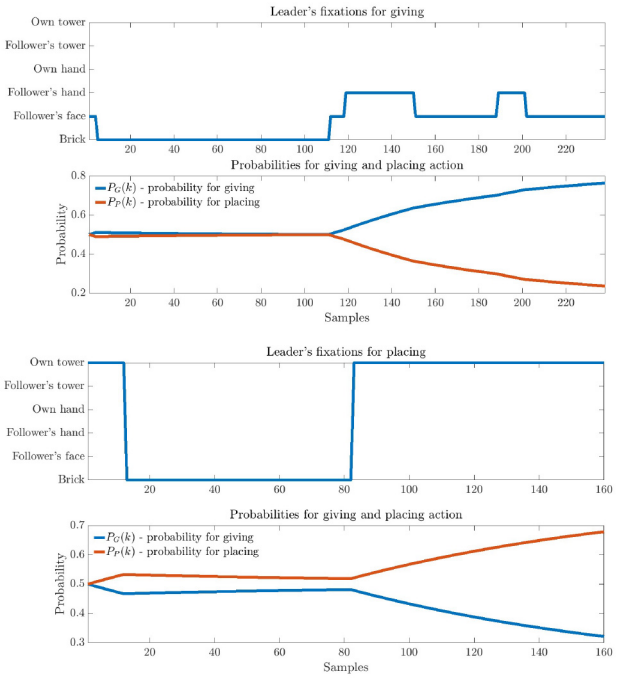

Fig. 9. Change of the signals $P_G(k)$ (blue line) and $P_P(k)$ (red line) with respect to the leader's gaze fixations for *giving* (two top figures) and *placing* action (two bottom figures).

correctly classified at around 60% completion which for an action that takes on average 2 s to finish, the system has a reaction time of 1.12 s. As for a *placing* action, it takes longer to predict, around 80% completion, which puts the reaction time at 1.36 s. The slower prediction could be caused by a prolong period of time fixating the brick, as shown in Fig. 3, which brings ambiguity to the system. In Fig. 9 there are two examples where, in the beginning, the Action Anticipation system cannot predict which of the actions is the leader performing. When the leader gaze fixation switches to the FF or hand, the probability for *giving* increases, and when the leader gaze fixations switch to his own tower, the probability for *placing* increases. The relation between the P_G and P_P signals is used to predict the leader's action, A^L_{k+1}.

The main conclusions from the modeling with human experiments are: 1) the *Gaze Dialogue Model* can generate gaze fixations for the *giving* and *placing* actions that are similar to the ones observed in the HHI data; 2) the *Gaze Dialogue Model* can predict accurately the follower's next gaze fixations, when provided with the leader's real gaze fixations, from the HHI dataset; and 3) it is possible to predict the correct actions from the gaze fixations using our *Gaze Dialogue Model*. The next section describes how to incorporate the model in a human–robot interaction scenario.

V. Human–Robot Interaction Experiment

This section addresses the validation of the *Gaze Dialogue Model* in an HRI scenario. We start by describing the two types of HRI: 1) robot-as-a-leader and 2) robot-as-a-follower; however, the focus of this work will be on robot-as-a-follower, for reasons that will become clearer below. Second, we describe the human-in-the-loop system, which is important for the

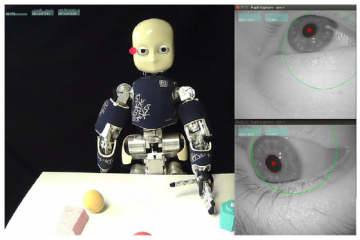

Fig. 10. Egocentric view of the human from the head-mounted eye-tracker with a red-marker indicating the current human fixation.

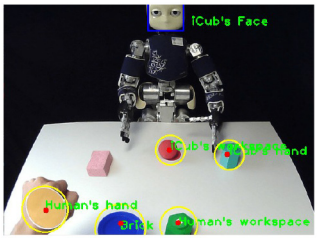

Fig. 12. All the important regions and correct labels are identified. Additional objects which are not relevant are considered outliers.

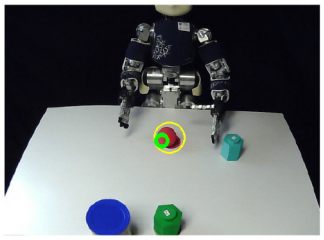

Fig. 11. Red ball detection is marked by the yellow circle, and the human gaze fixation is the green hallow circle.

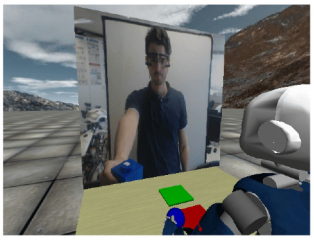

Fig. 13. Experimental setup for the HRI scenario. Human subject is wearing a head-mounted eye-tracker and the relevant objects are present.

interaction between robot-as-a-leader and robot-as-a-follower. The section finishes with a discussion on the results of HRI experiments and an analysis of the interaction in comparison to the HHI experiments.

Both HRI experimental scenarios, robot-as-a-leader and robot-as-a-follower, share three main common aspects. First, the human actor is equipped with the Pupil Labs eye tracker, introduced in Section III, to track the human gaze fixations while (s)he interacts with the robot. The software interface and the gaze fixation point are shown in Figs. 10–13. Second, concerning the low-level controllers, the eye-gaze saccadic movements in the iCub is driven by the Cartesian 6-DOF gaze controller described in [37]. As for the arm movements, a minimum jerk Cartesian controller is applied to control the iCub's arm and torso [38]. The iCub motor controllers run at 50 Hz. The HRI validation was made using the HHI dataset. The human switching from fixating the brick to another region-of-interest is usually associated with the beginning of either the *giving* or the *placing* action. Since in the robot-as-a-leader the leader is always aware of the action, it does not require any feedback from the follower. As such, the robot-as-a-leader scenario does not require the robot to sense any data from the human. In this work, we assume that once the robot takes the leader's role, the *Gaze Dialogue Model* generates the robot's gaze fixations and plans its action to execute either a *giving* or a *placing* action. The eye-gaze communication and arm movements are assumed to be communicated and 'read' by the human follower. The leader's eye-gaze communication for *giving* or *placing* actions is determined as the most likely gaze fixations observed in the HHI dataset. The robot-as-a-leader can be seen in the supplementary video material.

A. Robot Setup in the Leader–Follower Scenario

HRI experiments were carried out with the iCub robotic platform [39]. The iCub is a humanoid capable of performing actions that are "legible" to humans [24]. It has two cameras, on the head of the robot, that are capable of vergence and version, in a way similar to the human oculomotor control system.

In the case of the robot-as-a-follower, the human wears the eye-tracking system during *placing* and *giving* actions. As the robot has to follow the interaction, it has to interpret the relevant gaze fixations from the human. In this scenario, the human is part of the control loop, by providing feedback to the robot controller through his/her gaze fixations. This information is streamed, in real-time, to allow the robot to predict the human gaze fixations and actions while, also in real-time, generating the robot's own gaze fixations and planning the robot's own actions. Fig. 14 illustrates the human-in-the-loop modules involved in the HRI. In the next section, we will go through the added steps taken to integrate the human-in-the-loop in the robotic platform.

B. Human-in-the-Loop System

In the presented human-in-the-loop system, the human gaze fixations are provided to a robot. This is achieved from the eye-tracker data which composes of RGB image frames (from the eye-tracker world camera), and a 2-D pixel location of the *Gaze Fixation Point* that is segmented and labeled in the *Object Detection* and *Face Detection* systems. The segmentation and labeling procedure is inspired on the work from Samira and Odobez [40] which involves tracking region in the image fixated by the human observer. The following sections explain the necessary to segment, label, and communicate all the important eye-gaze fixations to the robot.

1) Gaze Fixation Point: The first step in the implementation of the diagram of Fig. 14 involves synchronizing the gaze fixation point provided by the LSL network [34], and the video frame received directly by the Capture software. The gaze fixation point is marked by a green hollow circle

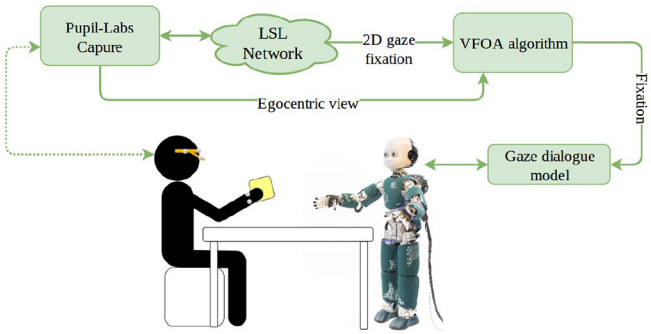

Fig. 14. Diagram illustrating the connections between the different modules that make up the communication of the human eye gaze to the robot fixations. The first module is related to the software that acquires the data from the eye tracker—Captured by Pupil Labs [33]. From this module, we collect the 2-D fixation point of the subject's gaze projected onto the world view camera on the eye tracker. The stream of the world view camera, together with 2-D gaze fixations through LSL network [34] (latency < 0.1 ms), is sent to the VFOA algorithm module to track the relevant fixations. The final module is the implementation of the *Gaze Dialogue* model described in Section IV.

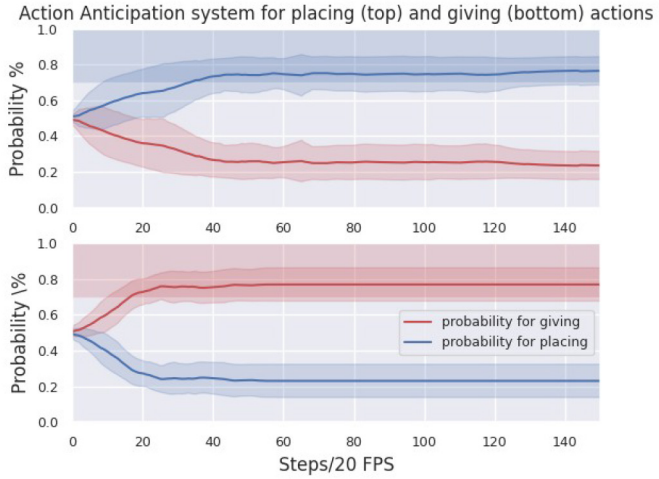

Fig. 15. Action anticipation results for HRI trials on classifying the actions as a placing or giving action, respectively.

in Fig. 11, and it is recorded at 120 Hz. The world camera, that is, the egocentric view of the human, is published at 60 frames per second (FPS). Since the frequency of the gaze fixation stream is two times faster than the stream of the world camera, we process every two gaze fixation points from the buffer sent by the LSL network. Whenever the green hollow circle, that is, the human eye-gaze fixation, is inside a region of interest, the visual focus of attention (VFOA) algorithm classifies the fixation point as a valid state S_k, and it is sent to the *Gaze Dialogue* model. The VFOA is inspired on the concept from [41] where it classifies important eye-gaze fixations the states S_k and, correspondingly, the observations O_k.

2) Object Detection: To classify the objects we use a color-based algorithm which extracts the relevant colors as the relevant objects to the HRI setup. The VFOA algorithm outputs the current object fixated if the *Gaze Fixation Point* is inside the region of one of the detected color objects. An example of an HRI setup with the VFOA algorithm classifying objects of different colors with its corresponding label is in Fig. 12. Table V identifies the objects, with the corresponding colors, extracted in the HRI experiments and the associated label given to the *Gaze Dialogue Model*.

TABLE V
ASSOCIATED LABEL TO THE COLORED OBJECT IN THE HRI SETUP

Label	color	RGB	Object
Brick	blue		Cylinder
iCub's Workspace	red		Sphere Shape
Human's Workspace	green		Hexagonal Shape
iCub's Hand	cyan		Cyan Sticker
Human's Hand	yellow		Yellow Sticker

3) Face Detection: For detecting the iCub's face, we apply a Haar cascade classifier algorithm. We created a new cascade trained with real images of the iCub's face. This classifier can detect the iCub's face in the HRI scenario quite accurately with very few false positives during the trials. Fig. 12 shows all the regions of interest, including the iCub's face, detected from the VFOA algorithm output.

The *Gaze Dialogue Model* was implemented in the Human-in-the-loop system as follows. First, the human eye-gaze fixations are used as observations O_k and the robot's gaze fixations as the current state S_k. Second, the robot can predict the leader's gaze fixation $\hat{S}^L_{k+1}$ and action $\hat{A}^L_{k+1}$, using the appropriate HMM in Table II and the Action Anticipation algorithm from Section IV-D. Third, the predictions are fed into the Planning/Control block. Fourth, The posterior decoding executes to generate the follower's gaze fixations S'^F_{k+1}. Finally, the leader's predicted action is used to plan the follower's action SA'^F_{k+1}. This information is used to determined which HMM model to apply in the iteration $k+1$ to generate the next eye-gaze communication and arm movement of both leader and follower. The follower's gaze fixations are given as input to the robot eye controller [37] to drive the eyes toward the correct 3-D space gaze fixation point. The Action Anticipation system decides whether the robot starts its arm movement toward the handover location, in the case of *giving*, or stand still, in the case of a *placing* action.

C. Results of the Human–Robot Interaction Experiments

Concerning the robot-as-a-follower experiments, we instruct the human to perform the two types of actions plus an additional one: 1) giving; 2) placing; and 3) fooling. The first two actions are the same used for the HHI experiment, hence the subject interacting with the robot, albeit naive to the previous experiment, acted naturally without any further instructions. A total of 40 trials with one participant, 20 trials performing both placing and giving actions. The human-in-the-loop system with the *Gaze Dialogue Model* ran online at 20 FPS and each step O_{k+1} is a labeled human gaze fixation. Fig. 15 shows the mean and standard deviation of the Action Anticipation systems for all of the trials in the HRI scenario. Most of the interactions are correctly classified (average of 75% or above) with 40 iterations which correspond to around 4–5 s of real-time human gaze fixation sequence.

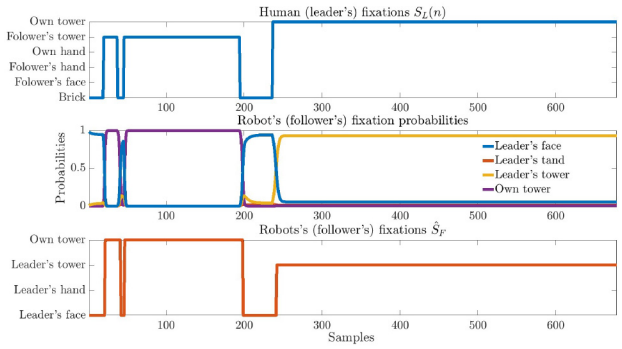

Fig. 16. Human and iCub's fixations when human as a leader is fooling a robot (starts with giving and after some time switches to placing action): leader's gaze fixations (top); probabilities of follower's fixations (middle); follower's decoded most likely gaze fixations (bottom).

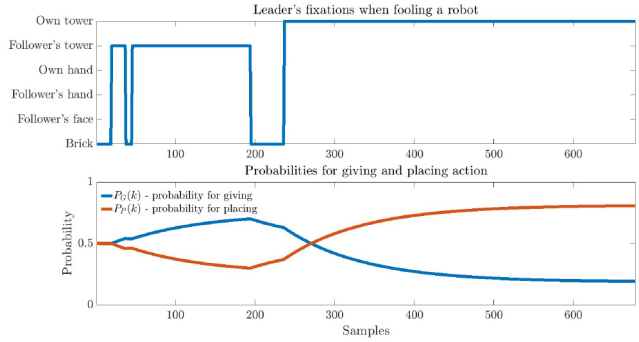

Fig. 17. Robot's action prediction when a human is fooling a robot (starts with giving and switches to placing action).

As for the third (fooling) action, the subject is instructed to cause a perturbation during the execution of a handover (*giving* action) by switching to a pick & place (*placing* action). The purpose is to show the active adaptation of the *Gaze Dialogue Model* to the different gaze fixations of the human. Fig. 16 shows the human gaze fixations. During the fooling action, the human begins handing over, and before completion, the human places the brick in her/his workspace. During the first 200 steps, the human gaze fixations are mostly on the follower's tower, which correlates with a *giving* action. After the 200th step, the human fixates his/her own tower, which is consistent with a *placing* action. According to its gaze fixations generated by the model, the robot initially fixates its tower, before successively fixating the LF and the LT. In short, the results of the gaze fixations for the fooling action illustrate a fast reaction to the nonverbal gaze communication cues exhibited at runtime. The *Gaze Dialogue Model* is capable of updating and revert the action classified.

In addition to the recorded gaze fixation probabilities, Fig. 17 shows the output of the Action Anticipation system and the predicted action of the human at each iteration. As the interaction starts, the *Gaze Dialogue Model* and, more specifically, the Action Anticipation system, predicts a *giving* action. The decision concerning the *giving* action was made when the difference between the signals $P_G(k)$ and $P_P(k)$ exceeds a predefined threshold. The threshold is empirically determined and it influences how fast the *Gaze Dialogue Model* reacts to nonverbal communication cues. This decision was used by the robot to decide whether the action is *giving*, as well as to start its arm movement, that is, arm reaching toward the handover location, or a *placing* action, to move the arm back to the rest position and continue observing. Once the leader fixates his/her tower, the probability for a *placing* action increases. As a result, the robot returns to its rest position while observing the human performing a *placing* action. This experiment validates the capability of the *Gaze Dialogue Model* to: 1) adapt to human gaze fixations; 2) update the action observed; 3) generate correct coupling robot-as-a-follower gaze fixations; and 4) plan the according action. All of this simultaneously and in real-time.

These tests lead us to the following conclusions: 1) the *Gaze Dialogue Model* is capable of generating gaze fixations, in the robot-as-a-follower, from human gaze fixations in real-time; 2) the gaze fixation sequence generated respect the human-like behavior observed in the HHI; 3) it can successfully predict the human action from gaze fixations in an HRI scenario; and 4) the human-in-the-loop system can translate online the human VFOA into relevant gaze fixations during the HRI experiments. Overall, the *Gaze Dialogue Model* learns from human eye-gaze cues the human action intention.

VI. Conclusion

The proposed *Gaze Dialogue Model* predicts the leader's gaze fixations and the action is inferred from the leader's gaze fixations. The posterior decoding is used to plan the follower's gaze fixations, based on previous fixations and the observed leader's gaze fixations. The inferred leader's action is used for both: 1) predicting the leader's gaze behavior and 2) posterior decoding of follower's gaze fixations.

Our contributions emerge during dyadic interactions involving individuals (*placing*) actions and actions-in-interaction (*giving*) actions. We implemented the model using the data collected during HHI experiments. The data consisted of paired, synchronized gaze fixations of people involved in the collaborative task. The *Gaze Dialogue Model* combines four HMMs that are selected based on the role of the person, leader or follower and two types of action: *giving* and *placing* for each role. The model was implemented in the iCub robot controller and tested in HRI scenarios. For the robot-as-a-leader scenario, the iCub produces nonverbal gaze communication signals (illustrated in the supplementary video.GazeDialogue.ieee-2022) that correlate with the instructed action and may thus be interpreted by the human. For the experiments in the robot-as-a-follower case, we use a human-in-the-loop approach, and the human gaze fixations are fed back to the model running in the robot controller. The human-in-the-loop allows the robot to: 1) infer the human action and 2) to adjust its gaze fixations according to the human action.

The iCub eye-gaze saccadic controller performed gaze fixations at a speed approximate to a human, which allows for an accurate representation of the *Gaze Dialogue Model* on the robot. On the other hand, due to hardware restrictions, the arm-motor movements were slower on average to a human. This delay between the eye-gaze fixations and the arm movements when performing actions resulted in longer execution times

when compared to the HHI experiments. Since the modeling of the robot's behavior is based on the HHI experiment data, if we have a robot with similar human arm-movements speeds, it is possible to achieve a more natural behavior of the robot. Another limitation is on the human–human social dynamic where the leader's goal is solely to concentrate on the task at hand without having to take the follower's actions into account. Therefore, the leader's action was considered to be known, and the model was deterministic, corresponding to the most likely gaze fixations of a human as a leader.

For future work, we will apply the *Gaze Dialogue Model* to more complex interactions, such as in collaborative assembly tasks where sequence of pick-and-place and handovers are present. This has the goal of exploring the advantage of including eye-gaze nonverbal communication in human–robot scenarios. We also want to evaluate the response time of the action prediction system in comparison to other approaches of human action prediction. The quantitative analysis of the human gaze fixations, as well as the reaction time are some of the metrics to evaluate the prediction capabilities of the model. The "mutual" action understanding is a field worth exploring in the context of social interaction and collaborative tasks for both human–human and human–robot contexts. Another step is to extend the model to handle new fixation points and/or missing/incomplete sequence of eye-gaze fixations.

We have stressed the importance of the *Gaze Dialogue* between two humans during an interaction, and how much information and coordination is achieved by this means. The *Gaze Dialogue Model* we proposed succeeds in partially capturing this interindividual coordination and, thus, provide a transparent mechanism that contributes to enhance the quality of interaction between humans and future generation of robots.

REFERENCES

[1] C.-M. Huang, S. Andrist, A. Sauppé, and B. Mutlu, "Using gaze patterns to predict task intent in collaboration," *Front. Psychol.*, vol. 6, p. 1049, Jul. 2015.

[2] J. Stanley and R. C. Miall, "Using predictive motor control processes in a cognitive task: Behavioral and neuroanatomical perspectives," *Progr. Motor Control*, vol. 629, pp. 337–354, 2009, doi: 10.1007/978-0-387-77064-2_17.

[3] N. Sebanz, H. Bekkering, and G. Knoblich, "Joint action: Bodies and minds moving together," *Trends Cogn. Sci.*, vol. 10, no. 2, pp. 70–76, 2006.

[4] P. Ricciardelli, E. Bricolo, S. M. Aglioti, and L. Chelazzi, "My eyes want to look where your eyes are looking: Exploring the tendency to imitate another individual's gaze," *Neuroreport*, vol. 13, no. 17, pp. 2259–2264, 2002.

[5] R. S. Johansson, G. Westling, A. Bäckström, and J. R. Flanagan, "Eye–hand coordination in object manipulation," *J. Neurosci.*, vol. 21, no. 17, pp. 6917–6932, 2001.

[6] N. Sebanz and G. Knoblich, "Prediction in joint action: What, when, and where," *Topics Cogn. Sci.*, vol. 1, no. 2, pp. 353–367, 2009.

[7] H. Yamamoto, A. Sato, and S. Itakura, "Eye tracking in an everyday environment reveals the interpersonal distance that affords infant-parent gaze communication," *Sci. Rep.*, vol. 9, no. 1, pp. 1–9, 2019.

[8] R. Kuboshita, T. Fujisawa, K. Makita, R. Kasaba, H. Okazawa, and A. Tomoda, "Intrinsic brain activity associated with eye gaze during mother–child interaction," *Sci. Rep.*, vol. 10, no. 1, 2020, Art. no. 18903.

[9] C. Bassetti, "Chapter 2—Social interaction in temporary gatherings: A sociological taxonomy of groups and crowds for computer vision practitioners," in *Group and Crowd Behavior for Computer Vision*, V. Murino, M. Cristani, S. Shah, and S. Savarese, Eds. New York, NY, USA: Academic, 2017, pp. 15–28.

[10] A. T. Duchowski, "Gaze-based interaction: A 30 year retrospective," *Comput. Graph.*, vol. 73, pp. 59–69, Jun. 2018.

[11] Z. Yäcel, A. A. Salah, Ç. Meriçli, T. Meriçli, R. Valenti, and T. Gevers, "Joint attention by gaze interpolation and saliency," *IEEE Trans. Cybern.*, vol. 43, no. 3, pp. 829–842, Jun. 2013.

[12] X. Yu et al., "Is interactional dissynchrony a clue to deception? Insights from automated analysis of nonverbal visual cues," *IEEE Trans. Cybern.*, vol. 45, no. 3, pp. 492–506, Mar. 2015.

[13] N. F. Duarte, M. Raković, and J. Santos-Victor, "Robot learning physical object properties from human visual cues: A novel approach to infer the fullness level in containers," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, 2022, pp. 10375–10381.

[14] P. Schydlo, M. Raković, L. Jamone, and J. Santos-Victor, "Anticipation in human–robot cooperation: A recurrent neural network approach for multiple action sequences prediction," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2018, pp. 5909–5914.

[15] S. Ivaldi, S. M. Anzalone, W. Rousseau, O. Sigaud, and M. Chetouani, "Robot initiative in a team learning task increases the rhythm of interaction but not the perceived engagement," *Front. Neurorobot.*, vol. 8, p. 5, Feb. 2014.

[16] R. T. Chadalavada, H. Andreasson, M. Schindler, R. Palm, and A. J. Lilienthal, "Bi-directional navigation intent communication using spatial augmented reality and eye-tracking glasses for improved safety in human–robot interaction," *Robot. Comput. Integr. Manuf.*, vol. 61, Feb. 2020, Art. no. 101830.

[17] A. Pereira, C. Oertel, L. Fermoselle, J. Mendelson, and J. Gustafson, "Effects of different interaction contexts when evaluating gaze models in HRI," in *Proc. ACM/IEEE Int. Conf. Human–Robot Interaction (HRI)*, 2020, pp. 131–139.

[18] K. Kompatsiari, F. Ciardo, V. Tikhanoff, G. Metta, and A. Wykowska, "It's in the eyes: The engaging role of eye contact in hri," *Int. J. Soc. Robot.*, vol. 13, no. 3, pp. 525–535, 2021.

[19] G. Perugia, M. Paetzel-Prüsmann, M. Alanenpää, and G. Castellano, "I can see it in your eyes: Gaze as an implicit cue of uncanniness and task performance in repeated interactions with robots," *Front. Robot. AI*, vol. 8, p. 78, Apr. 2021.

[20] O. Palinko, F. Rea, G. Sandini, and A. Sciutti, "Eye gaze tracking for a humanoid robot," in *Proc. IEEE-RAS 15th Int. Conf. Humanoid Robots (Humanoids)*, Nov. 2015, pp. 318–324.

[21] J. Domhof, A. Chandarr, M. Rudinac, and P. Jonker, "Multimodal joint visual attention model for natural human–robot interaction in domestic environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2015, pp. 2406–2412.

[22] L. Shi, C. Copot, and S. Vanlanduit, "GazeEMD: Detecting visual intention in gaze-based human–robot interaction," *Robotics*, vol. 10, no. 2, p. 68, 2021.

[23] L. Chukoskie et al., "Quantifying gaze behavior during real-world interactions using automated object, face, and fixation detection," *IEEE Trans. Cogn. Develop. Syst.*, vol. 10, no. 4, pp. 1143–1152, Dec. 2018.

[24] N. F. Duarte, M. Raković, J. Tasevski, M. I. Coco, A. Billard, and J. Santos-Victor, "Action anticipation: Reading the intentions of humans and robots," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 4132–4139, Oct. 2018.

[25] M. Raković, N. Duarte, J. Tasevski, J. Santos-Victor, and B. Borovac, "A dataset of head and eye gaze during dyadic interaction task for modeling robot gaze behavior," in *Proc. MATEC Web Conf.*, vol. 161, 2018, Art. no. 03002.

[26] S. Andrist, M. Gleicher, and B. Mutlu, "Looking coordinated: Bidirectional gaze mechanisms for collaborative interaction with virtual characters," in *Proc. ACM CHI Conf. Human Factors Comput. Syst.*, New York, NY, USA, 2017, pp. 2571–2582.

[27] M. Otsuki, K. Maruyama, H. Kuzuoka, and Y. Suzuki, "Effects of enhanced gaze presentation on gaze leading in remote collaborative physical tasks," in *Proc. CHI Conf. Human Factors Comput. Syst. (CHI)*, 2018, p. 368.

[28] S. Grogorick, M. Stengel, E. Eisemann, and M. Magnor, "Subtle gaze guidance for immersive environments," in *Proc. ACM Symp. Appl. Percept. (SAP)*, 2017, p. 4.

[29] L. Lukic, J. Santos-Victor, and A. Billard, "Learning robotic eye–arm–hand coordination from human demonstration: A coupled dynamical systems approach," *Biol. Cybern.*, vol. 108, no. 2, pp. 223–248, 2014.

[30] J. Ondras, O. Celiktutan, P. Bremner, and H. Gunes, "Audio-driven robot upper-body motion synthesis," *IEEE Trans. Cybern.*, vol. 51, no. 11, pp. 5445–5454, Nov. 2021.

[31] P. Kratzer, S. Bihlmaier, N. B. Midlagajni, R. Prakash, M. Toussaint, and J. Mainprice, "Mogaze: A dataset of full-body motions that includes workspace geometry and eye-gaze," 2020, *arxiv.abs/2011.11552*.

[32] M. Gallotti, M. Fairhurst, and C. Frith, "Alignment in social interactions," *Consciousness Cogn.*, vol. 48, pp. 253–261, Feb. 2018.
[33] M. Kassner, W. Patera, and A. Bulling, "PUPIL: An open source platform for pervasive eye tracking and mobile gaze-based interaction," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. Adjunct Publication*, 2014, pp. 1151–1160.
[34] C. Kothe. "Lab Streaming Layer (LSL)." Feb. 2018. [Online]. Available: https://github.com/sccn/labstreaminglayer
[35] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.
[36] Q.-L. Li, "Markov reward processes," in *Constructive Computation in Stochastic Models With Applications*. Berlin, Germany: Springer, 2010, pp. 526–573.
[37] A. Roncone, U. Pattacini, G. Metta, and L. Natale, "A Cartesian 6-DoF gaze controller for humanoid robots," in *Proc. Robot. Sci. Syst.*, 2016, pp. 1–9.
[38] U. Pattacini, F. Nori, L. Natale, G. Metta, and G. Sandini, "An experimental evaluation of a novel minimum-jerk Cartesian controller for humanoid robots," in *Proc. IEEE Intell. Robots Syst. (IROS) IEEE/RSJ Int. Conf.*, 2010, pp. 1668–1674.
[39] G. Metta et al., "The icub humanoid robot: An open-systems platform for research in cognitive development," *Neural Netw.*, vol. 23, nos. 8–9, pp. 1125–1134, 2010.
[40] S. Sheikhi and J.-M. Odobez, "Recognizing the visual focus of attention for human robot interaction," in *Human Behavior Understanding*, A. A. Salah, J. Ruiz-del Solar, Ç. Meriçli, and P.-Y. Oudeyer, Eds. Berlin, Germany: Springer, 2012, pp. 99–112.
[41] D. Das, M. G. Rashed, Y. Kobayashi, and Y. Kuno, "Supporting human–robot interaction based on the level of visual focus of attention," *IEEE Trans. Human–Mach. Syst.*, vol. 45, no. 6, pp. 664–675, Dec. 2015.

Mirko Raković (Member, IEEE) received the M.Sc. and Ph.D. degrees from the University of Novi Sad, Novi Sad, Serbia, in 2005 and 2013, respectively.

He is an Associate Professor of Robotics and Mechatronics with the University of Novi Sad, where he leads the Laboratory for Robotics and Mechatronics, Faculty of Technical Sciences. He was a Postdoctoral Researcher with the Institute Superior Técnico, University of Lisbon, Lisbon, Portugal, from 2017 to 2018, as part of Vislab Laboratory. His research interests include robotics and artificial intelligence, specifically HRI, biped locomotion, mechatronics, and applications of AI to robotic systems.

Nuno Ferreira Duarte (Graduate Student Member, IEEE) received the B.Sc./M.Sc. degree in electrical and computer engineering with specialization in systems, decision and control from the Instituto Superior Técnico, Lisbon, Portugal, in September 2016. He is currently pursuing the joint Ph.D. degree with the Instituto Superior Técnico and the Swiss Institute of Technology Lausanne (EPFL), Lausanne, Switzerland, advised by Prof. J. Santos-Victor and Prof. A. Billard.

He previously interned with the California Institute of Technology, Pasadena, CA, USA, under the supervision of Prof. R. Murray. His research interests include HRI, machine learning, and robotics.

Jorge Marques received the Ph.D. and aggregation degrees from the Technical University of Lisbon, Lisbon, Portugal, in 1990 and 2002, respectively.

He is currently a Professor with the Electrical and Computer Engineering Department, Instituto Superior Técnico, Lisbon, and a Researcher with the Institute for Systems and Robotics. His research interests are in image processing, shape analysis, and pattern recognition.

Prof. Marques was the President of the Portuguese Association for *Pattern Recognition* and an Associate Editor of the *Statistics and Computing* (Springer).

Aude Billard (Fellow, IEEE) received the B.Sc. and M.Sc. degrees in physics from the Swiss Institute of Technology Lausanne (EPFL), Lausanne, Switzerland, in 1994 and 1995, respectively, and the Ph.D. degree in artificial intelligence from the University of Edinburgh, Edinburgh, U.K., in 1998.

She is a Full Professor and the Head of the LASA Laboratory, School of Engineering, EPFL. Her research spans the fields of machine learning and robotics with a particular emphasis on learning from sparse data and performing fast and robust retrieval.

Prof. Billard was the recipient of the IEEE-RAS Best Reviewer Award and IEEE-RAS Distinguished Service Award. Her research received the Best Paper Awards from the IEEE TRANSACTIONS ON ROBOTICS, RSS, ICRA, and IROS. She is the President-Elect of the IEEE Robotics and Automation Society (RAS) and a former IEEE RAS Vice-President of publication activities.

José Santos-Victor (Fellow, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in electrical and computer engineering from Instituto Superior Técnico, Lisbon, Portugal, in 1988, 1991, and 1995, respectively.

He is a Full Professor with the Department of Electrical and Computer Engineering, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, where he is the President of the Institute for Systems and Robotics. He conducts multidisciplinary research, with neuroscientists and psychologists, in the areas of cognitive and bioinspired computer vision and robotics, including the design of advanced humanoid robotic platforms like the iCub.

Prof. Santos-Victor is a member of the Academy of Sciences of Lisbon.