Action Alignment from Gaze Cues in Human-Human and Human-Robot Interaction

Nuno Ferreira Duarte[1], Mirko Raković[1,2], Jorge Marques[1] and José Santos-Victor[1] ⋆

[1] Vislab, Institute for Systems and Robotics, Instituto Superior Técnico, Universidade de Lisboa, Portugal
{nferreiraduarte, rakovicm, jsm, jasv}@isr.tecnico.ulisboa.pt
[2] Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia
rakovicm@uns.ac.rs

Abstract. Cognitive neuroscience experiments show how people intensify the exchange of non-verbal cues when they work on a joint task towards a common goal. When individuals share their intentions, it creates a social interaction that drives the mutual alignment of their actions and behavior. To understand the intentions of others, we strongly rely on the gaze cues. According to the role each person plays in the interaction, the resulting alignment of the body and gaze movements will be different. This mechanism is key to understand and model the socially dyadic interactions.

We focus on the alignment of the leader's behavior during dyadic interactions. The recorded gaze movements of dyads are used to build a model of the leader's gaze behavior. The use of the follower's gaze behavior data is two-fold: (i) to determine whether the follower is involved in the interaction, and (ii) if the follower's gaze behavior correlates to the type of the action under execution. This information is then used to plan the leader's actions in order to sustain the leader/follower alignment in the social interaction.

The model of the leader's gaze behavior and the alignment of the intentions is evaluated in a human-robot interaction scenario, with the robot acting as a leader and the human as a follower. During the interaction, the robot (i) emits non-verbal cues consistent with the action performed; (ii) predicts the human actions, and (iii) aligns its motion according to the human behavior.

Keywords: Action Anticipation, Gaze Behavior, Action Alignment, Human-Robot Interaction

⋆ Manuscript received May 31, 2018; Corresponding author: Nuno Duarte. Work supported by EU H2020 project 752611 - ACTICIPATE, FCT project UID/EEA/50009/2013 and RBCog-Lab research infrastructure. We thank all of our colleagues, students and volunteers that supported us in preparing and conducting the experiments.

1 Introduction

Humans can interact with the environment, objects, or with other humans. Interacting with the environment and objects requires visually adjusting our movements in order to correctly perform the intended action. The interaction with other humans requires the contribution of different components. Humans use verbal communication to express motion and intent to others. However, since verbalizing every step of the interaction would be time-consuming and cognitively expensive, humans use the body as a communication tool. This means that while we are executing our intended action, we are also communicating to others the exact action we are performing. This capacity is referred to as non-verbal communication and involves all the motion degrees of freedom in our bodies: from pointing a finger expressing a direction of interest, to a saccadic eye movement to specify a place that attracted our attention.

The work described in [6] investigates how the non-verbal communication cues of one human allows the others to read his action intentions. The non-verbal communication of the actor was recorded using a motion tracking system for the motion of the body, and a head mounted eye tracker for the gaze behavior of the eyes. The scenario involved one actor, interacting with 3 humans, and performing one of two actions: *placing* of an object on a table, or *giving* the object to one of the humans facing him. These actions were chosen as they fall into two categories of actions defined in micro-sociological studies [3]. The *placing* action is an instance of an *individual action*, while the *giving* action is part of the category *action-in-interaction*, that requires for communications between the interaction partners.

The focus of [6] was on the importance of the different non-verbal communication cues: arm movement, head movement, and eye movement. A human study was performed in which subjects watched short fragments of videos of the actor performing one of two possible actions. These fragments contain different amounts of information concerning the non-verbal cues, and the objective was to analyze the impact of each cue on the capacity to "read" the intentions of the actor. The data collected was used to model the arm behavior for the two types of actions, and to propose a gaze controller that, combined with the arm movement, is able to generate human-like movements, just like those observed in the Human-human interaction (HHI) experiments. This was corroborated by building a robotic controller that, when applied to a humanoid robot to perform the same actions, allows human subjects to understand the robot's intentions from the video fragments, with an accuracy similar to the case of a human actor.

Nevertheless, the work was incomplete as it only studied the behavior of one of the parts of the interaction. So the logical step was to study not only the non-verbal communication of the human performing the action, but also the communication cues emitted by the second participant in the interaction. The focal point of Raković et. al. paper [23] was on the eyes' non-verbal communication, and the "gaze dialogue" model derived to couple the agent's gaze behavior. Each agent's behavior was modeled as a Hidden Markov Model (HMM), where the states were the gaze fixations, and the observations the gaze fixations of the

other agent. However, the approach discusses the prediction of one agent's action from his gaze fixations in order to adapt the gaze behavior of the second agent for an improved collaboration.

We adopt the terminology of [10] concerning the interaction roles, where one agent can be viewed as the leader and the other one as the follower, in the sense that the follower adapts his/her behavior to the leader, but not the other way around. Hence, in a human-robot interaction (HRI) scenario, a robotic follower will adapt to a human leader. However, when the robot is the leader, the model behaves deterministically and it does not adapt to the behavior of the human follower. In this case, the robot (leader) does not take the speed of the human participant into account, and it is not concerned with the human's understanding of the action. The contribution of the current paper is on tackling this issue.

In [23] the leader's gaze behavior was pre-defined as the average, most likely behavior observed from the HHI scenario. Although this behavior may work on average for most interactions, an HRI is never deterministic since humans are naturally unpredictable and stochastic. As such, a reliable model for the leader's behavior needs to take the feedback of the follower's behavior into account. In this way, it becomes possible to achieve the third level of interaction [10], where both agents, the leader and the follower, adapt to each other in order to achieve a mutual alignment. The focus of this work is on closing the loop of the mutual alignment, by adapting the behavior of the actor performing the action (leader), to the behavior of the actor observing and eventually participating in the interaction (follower).

Section 2 discusses the relevant work done in the quest of understanding non-verbal communication, as well as on human action anticipation, when humans interact with other humans or objects. Section 3 describes the dataset and the HHI scenario used in this work, and the analysis of the data collected from the head mounted gaze tracker. The modeling of the gaze behavior is included in Section 4 and the HRI implementation with the results are shown in Section 5. The paper ends with a discussion of the results obtained, followed by an overall conclusion and delineating future work challenges.

2 Related Work

HRI requires the human and the robot to understand each other [27]. Modeling the interaction between agents has been tackled in several fields, including robotics, computer vision, and cognitive and behavioral science. Lukic et al. [18] presented the intrapersonal model for manipulating objects based on Gaussian Mixture Models to generate human-like behavior of the hand, arm, and eyes. This was later adapted to human-robot interaction in [6] to yield human-like behavior when involving non-verbal communication. Furthermore, the model was adapted in [23] to describe the non-verbal cues of the eyes of two agents using a cross-agent HMM.

There have been other approaches for modeling the eye gaze behavior over the years [7]. S. Ivaldi et al. [12] developed a robotic controller that uses the

head gaze orientation to understand which object the human is gazing at. One drawback is the use of head orientation as a proxy to estimate the eye gaze. In [5], the eye gaze estimates are used to understand the fixation point of humans. This combines eye tracking data with pointing gestures extracted from RGB-depth cameras, to estimate eye gaze fixation. The limitation with this approach is that all the processing is done off-line, and not during the interaction. Andrist et. al. [2] studied the gaze interaction of a human with a virtual agent in a sandwich-making task based on HHI experiments to improve the speed of the collaboration. However, this work only applies to the 'instructor role', that we designate as the leader's perspective, and lacks generality.

Palinko et al. [20] identify the pupil position in the eye in order to estimate the gaze direction. Despite not requiring any additional hardware to track the gaze orientation, they are constrained by the limited resolution of the iCub robot cameras and the accuracy will depend on lighting conditions. As for detecting joint attention, [28] describes work on the extraction of the gaze direction from the head pose of the human. Instead, we intend to extract the visual information collected with the two eye trackers during the HHI experiment scenario, that is publicly available from the Raković et. al. [24].

Regarding action anticipation, there has been research on the understanding of human motion [15], modeling the human motion to infer the executed action [29] and predicting human trajectories to trace a path of least collision for the robot, [22]. The prediction algorithm takes into account the human-environment and human-human natural adaptation to calculate the optimal path for the robot. Farhan et. al. [8] instead focus on predicting the action happening in the long future, instead of anticipating the ongoing action, using pre-recorded videos trained in large datasets of humans performing several different actions.

Koppula et. al. [16] include a rich dataset of human poses and objects to classify the action. However, it does not take advantage of the gaze behavior of humans to predict the action sooner and with higher accuracy. There are several papers presenting the use of human body coordinates, and only very few have gaze information, often limited to a couple of example scenarios [1, 9, 11].

Schydlo et al. [26] developed a learning based action anticipation model using motion and gaze fixation data of the human-human interaction experiment from the publicly available dataset of [6]. The model can quite accurately perform an early anticipation of the ongoing action, using a combination of the body and gaze coordinates. This action anticipation model uses a recurrent neural network to learn the non-verbal cues that the body and gaze behavior provide in order to distinguish between two actions: a *giving* or *placing* action. Although it can accurately predict the action at an early stage, the information given to the network can not be generalized to different HHI or HRI scenarios. Additionally, it does not provide the robot with any information on how to behave after the action is predicted, thus breaking any possibility of mutual understanding and alignment. Moreover, the results in [26] were deterministic, meaning it would give the same output when given the same data. Instead, the human behavior is stochastic and mutual alignment requires the robot to adapt to a specific

participant and not to an average behavior of a group of humans. In this paper, we discuss the importance of the two agents aligning with each other, and an approach where the agents exchange information from each other in order to predict the other's action, and adapt his/her own behavior.

3 Dyad Interaction Experiment

The dyad interaction experiment is composed of two actors participating in a joint task (Fig. 1). The two actors have to perform a turn-taking task of *placing* an object on the table, or *giving* the object to the other person. From this experiment we collect the gaze fixations of 6 participants, i.e. 3 dyads. We get a total of 72 actions seen from two perspectives. Out of 72 actions, 36 actions were *giving* and 36 were *placing*. The gaze fixations are tracked using the Pupil Labs eye tracker [13]. These sensors are connected through an LSL Network [17] which synchronizes and collects the data together with cameras recording the interaction - the egocentric view camera gives the subject's perspective. The gaze behavior of all 144 actions are labeled with identified relevant fixations and events throughout the action. The fixations are object (i.e. brick), team-mates' face (TM face), team-mates' hand (TM hand), own hand, team-mates' tower (TM tower), and own tower; and the events are object picked, object handed over, and object placed. Object handed over exists only in the *giving* action. In [24] it can be found a detail description of the experimental set-up and the data acquisition procedure. The focus of this paper is two-fold: (i) the gaze behavior of the leader during the *giving* action, more specifically on how he/she behaves before and after the handover, and (ii) follower's gaze fixation behavior when the action is *giving* or *placing*.

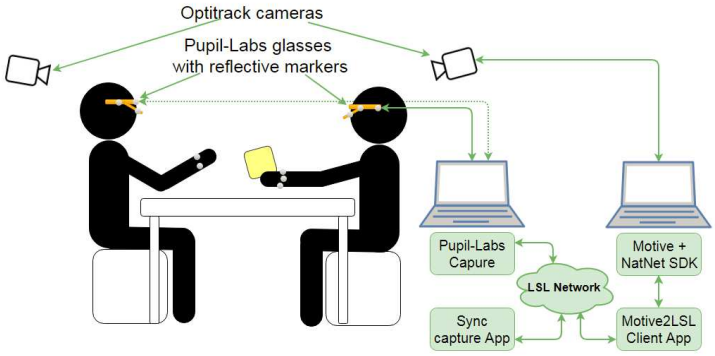

Fig. 1. Representation of the HHI experimental set-up and all the different communication systems. The image is taken from [24].

Fig. 2 shows the time spent on each of these gaze fixation states, throughout the whole action, and for the two perspectives. In addition to the total amount

of time spent on each state, we distinguish the gaze behavior before and after the handover. For these experiments, the handover time is defined as the moment when the leader's hand releases the object, and it is identified by the change in the fingers acceleration with respect to the brick.

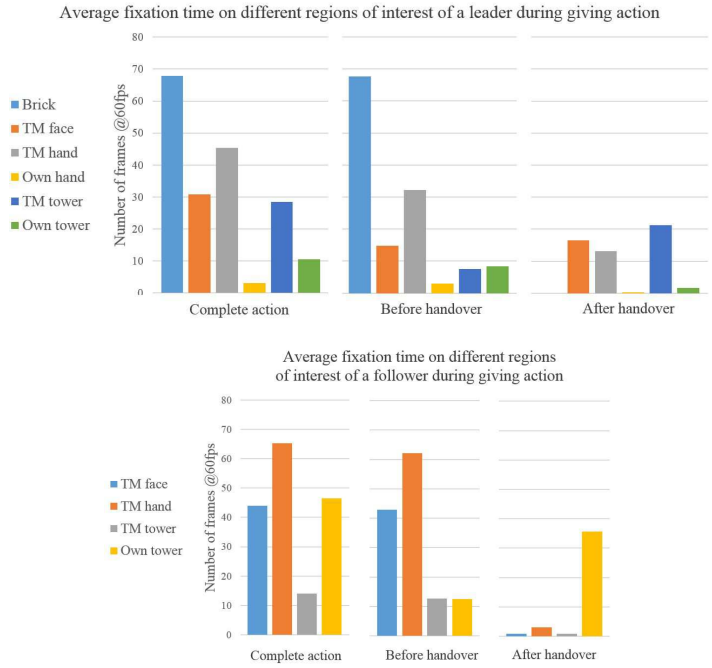

Fig. 2. Cumulative analysis of the gaze behavior during the HHI experiment for the complete action, before and after handover, showing the leader's (top) and the follower's fixations (bottom).

Fig. 2 (top image) shows how the leader is mainly focused at the object, and the TM face and hand, right before the handover. The brick is fixated when the leader is visual searching and/or grasping the object - the gaze assisting the motor control function. After the object is grasped, the leader looks mainly at the TM face, hand, and towers - the non-verbal cues to communicate the intention - the gaze engaged in communication purposes. Before the handover, Fig. 2 (bottom image), the follower fixates the TM's face and hand, aiming at reading the action intention of the leader - communicative gaze. After the handover, the non-verbal cues serve purely functional goals. As the object is already in the follower's possession, the remainder of the action requires the follower to fixate his own tower and controlling the arm towards the goal - the functional role of gaze to assist the motor control.

In the next section, the information from the HHI dataset is used to model the leader's behavior. The leader's gaze data will be used to model the stochastic behavior of the human that is different before and after the handover. The follower's gaze behavior will be used to retrieve his/her own understanding of the action, which is then provided to the leader to assess the follower's engagement in the interaction.

4 Modeling of the leader's behavior

Fig. 3 shows the block diagram for modeling the gaze behavior and aligned motion planning of agents P_1 and P_2. The state of each agent is defined as the gaze fixation S_k and type of action A_k. The fixations $[S_1(k), S_1(k-1), ...]$ are emitted by agent P_1, which are from the perspective of agent P_2, represented as observations $[O_1(k), O_1(k-1), ...]$. Simultaneously, fixations $[S_2(k), S_2(k-1), ...]$ are emitted by agent P_2, and represented as observations $[O_2(k), O_2(k-1), ...]$ of agent P_1.

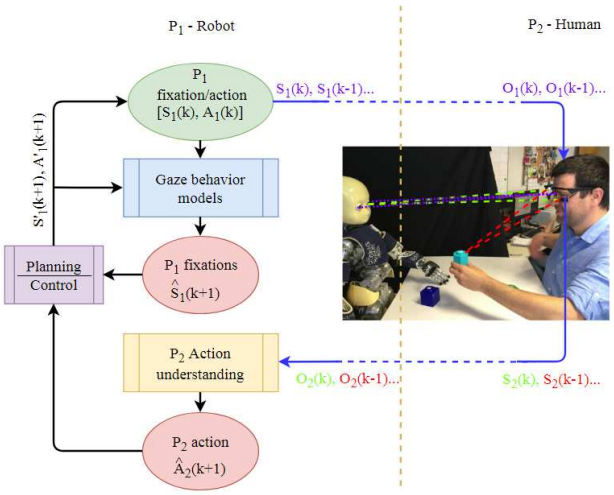

Fig. 3. Block diagram of the proposed leader's gaze behavior and alignment model. Agent P_1 emits fixations S_1 which corresponds to a particular action A_1. From the 'Gaze behavior models' it is generated the next fixation, $\hat{S}_1(k+1)$, from the previous knowledge, $S_1(k)$ and $A_1(k)$. The $\hat{S}_1(k+1)$ is the next fixation without the influence of agent P_2 in the interaction, i.e. without mutual alignment. Agent P_1 observation, O_2(k), is used to calculate the understanding of agent P_2, $\hat{A}_2$(k+1). This is then fed to the 'Planning/Control' block, together with the next fixation $\hat{S}_1(k+1)$, to estimate the new fixation and action of agent P_1, $S_1^{'}$(k+1) and $A_1^{'}$(k+1), respectively.

The central parts on Fig. 3 correspond to the gaze behavior models (blue block) and human action understanding (yellow block) and will be detailed in

Sections 4.1 and 4.2, respectively. The 'Gaze behavior models' encode the leader's gaze stochastic behavior, that depends on the type of action (in this paper the focus is on modeling the *giving* action, i.e. action-in-interaction) and can change over time after a significant event (i.e. object handover). Action understanding uses the gaze fixation of the human to estimate the probabilities of *giving* versus *placing* action. This is fed back to the 'Planning/Control' block for the motion planning of the agent and selection of appropriate gaze behavior model.

4.1 Gaze behavior of the leader

The leader's gaze behavior is modeled with Discrete-Time Markov Chains (DTMC) [4]. A DTMC represents the evolution of a system that stochastically switches from one state to another, at discrete time instances. The model has an associated internal state variable: $S_k \in \{U_1, ..., U_N\}$ where $U_1, ..., U_N$ denotes admissible state values, i.e. fixations, and $k \in \{1, ..., T\}$ denotes the discrete time instants. In the case of a *giving* action, the leader has six admissible states before the handover, and four states after (Fig. 4). This corresponds to the top image from Fig. 2 with six fixations before handover. After the handover, the brick is never fixated and the fixation of one's own hand is negligibly small.

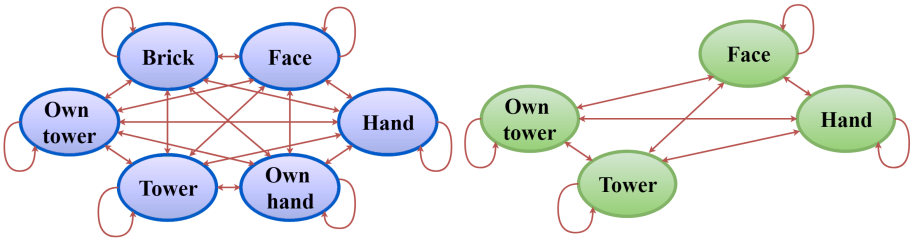

Fig. 4. DTMC for the behavior of a leader: (left) before the brick handover; (right) after the brick handover.

The two DTMCs (for the period before and after the handover) are represented by transition matrices learned from the HHI data, which has labeled fixations of the dyad throughout all the actions. Transitions of the fixations for *giving* before and after handover are counted, and the obtained transition matrices are given in Table 1.

The admissible states that correspond to the indexes of the rows and columns of the transition matrices are: 1 - Brick, 2 - TM Face, 3 - TM Hand, 4 - Own hand, 5 - TM tower and 6 - Own tower, before handover; and 1 - TM Face, 2 - TM Hand, 3 - TM tower and 4 - Own tower, after handover. To illustrate the output behavior that can be obtained with the DTMCs, we generated the fixation sequence of 400 samples (Fig. 5), the first 200 samples using the DTMC before handover and 200 samples using the DTMC after handover. Fig. 5 show that the fixations before handover are the brick, follower's face, and hand. After

Table 1. Transition matrix before handover A^L_{bhon} and after handover A^L_{ahon} for the *giving* action

Handover	Leader
Before	$A^L_{bhon}=\begin{bmatrix} 0.9861 & 0.0016 & 0.0045 & 0.0016 & 0.0041 & 0.0020 \\ 0.0038 & 0.9505 & 0.0438 & 0.0019 & 0 & 0 \\ 0.0018 & 0.0211 & 0.9718 & 8.81e^{-04} & 0.0044 & 0 \\ 0 & 0 & 0.0571 & 0.933 & 0.0095 & 0 \\ 0.0072 & 0.0145 & 0.0435 & 0.0036 & 0.9239 & 0.0072 \\ 0.0566 & 0.0031 & 0.0031 & 0.0126 & 0 & 0.9245 \end{bmatrix}$
After	$A^L_{ahon}=\begin{bmatrix} 0.9623 & 0.0205 & 0.0154 & 0.0017 \\ 0.0309 & 0.9423 & 0.0247 & 0.0021 \\ 0.0196 & 0.0039 & 0.9712 & 0.0052 \\ 0.0179 & 0.0179 & 0 & 0.9643 \end{bmatrix}$

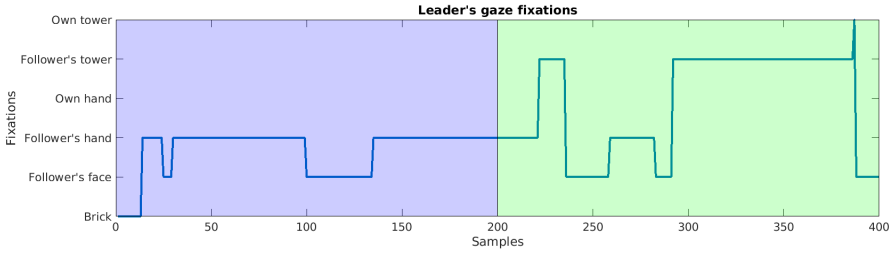

Fig. 5. Leader's fixations when is applied the DTMC before handover (blue section) and DTMC after handover (green section).

the handover, the fixations are the follower's face, hand, and tower, with very short fixation of the own tower. The leader's fixation are given in the top image of Fig. 2.

4.2 Human action understanding

Referring to Fig. 3, the robot (agent P_1) has access to the fixations of the human (agent P_2) which are represented as observations $O_2(k) \in \{V_1, ..., V_M\}$. The admissible fixations of the human are denoted by $V_1, ..., V_M$. The type of action is inferred from the HHI data of the follower's gaze fixations, by calculating the (average) empirical probabilities for *giving* versus *placing* conditioned to the follower's fixation, see Table 2.

When the follower looks at the leader's face, the probabilities for *giving* and *placing* are respectively 49.5% and 50.5%, meaning that it is not a strong cue for the action. Instead, when the follower looks at the leader's hand or at his own tower, it signals that the follower understood that the leader intends to give him

Table 2. Average probabilities for the *giving* and *placing* actions, with respect to the follower's gaze fixations

	Giving	Placing
Leader's face	0.495	0.505
Leader's hand	0.617	0.383
Leader's tower	0.293	0.706
Own tower	0.844	0.156

the brick. Finally, if the follower fixates the leader's tower, this is a strong signal that the follower understood that the leader will perform a *placing* action.

To select which action is being performed, we estimate the an action probability by combining the information related to the instantaneous follower's fixations, with the past history of that probability. These probability signals are denoted as P_G and P_P, respectively for the *giving* and *placing* actions.

Based on the current instantaneous follower's fixation, we use the action probabilities from Table 2, to update P_G) and P_P with an exponential moving average:

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

where k refers to time, and $\alpha = 0.05$. The update $\delta(k)$ depends on the values of Table 2, evaluated with the instantaneous follower's fixations. If the follower is currently fixating the leader's hand, and the *giving* action is selected, P_G is updated with $\delta(k) = 0.617$, and P_P is updated with $\delta(k) = -0.617$. If the *placing* action is selected, P_G is updated with $\delta(k) = -0.383$, and P_P is updated with $\delta(k) = 0.383$. This mechanism ensures a smooth evolution of the action probabilities and filters out spurious noisy measurements.

An example of human fixation, and the output of action understanding block are given in Figs. 7 and 9. In Fig. 7, the human is engaged in the action and the probability of *giving* is always higher than the probability for *placing*. However, in the second example, during a certain period of time, the human fixates the leader's tower, communicating that he is understanding that the agent will perform a *placing* action. In this period, the probability for *placing* grows, until the human switches the fixations to the agent's hand or its own tower. The second example will illustrate on-line alignment of the leader's action planning from the follower's gaze cues.

5 Human Robot Interaction Experiment

We used the iCub robotic platform [19] for our experiments. As a humanoid robot, the iCub has a body structure that is similar to the human body, so that humans can more easily understand the robot's motor behavior and, hence, its intentions [6], [14]. The eyes of the robot are 2 cameras capable of vergence and version movements, as in the human oculomotor system.

We used the same HRI experiment scheme as in [23], with the objective to track the gaze fixations of the human as a follower, while (s)he interacts with the robot. The gaze fixations are tracked with the Pupil Labs tracker, see Section 3.

A Cartesian-based gaze controller [25] was used to control the robot's eyes when fixating 3D coordinate points. The motor control of the torso, arm, hand, and fingers was done with a minimum jerk Cartesian controller [21], which is responsible for guiding the movement of the robot to grasp the object, as well as to move the object to the handover location, and return to the resting position.

Fig. 6 shows a robot performing a *giving* action. The HRI experiment starts with the human not attending to the robot, and looking at his notebook. During

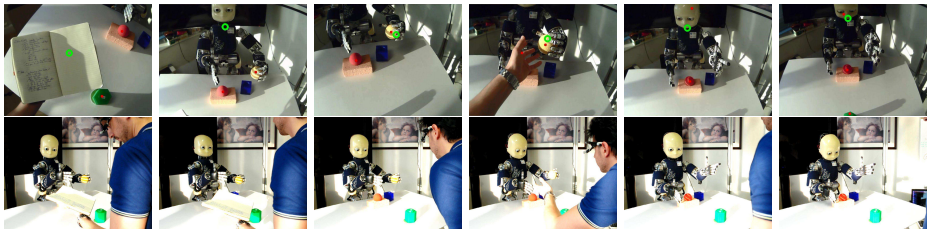

Fig. 6. The first experiment of a robot interacting with a human initially disengaged from the interaction. The green hallow circle in the top row images is the human gaze fixation. The red dots mark the important interaction cues (robot's face, robot's hand, robot's tower, own tower). When the green circle is in the region of interest of the red dot, then it is classified as the human looking at that cue.

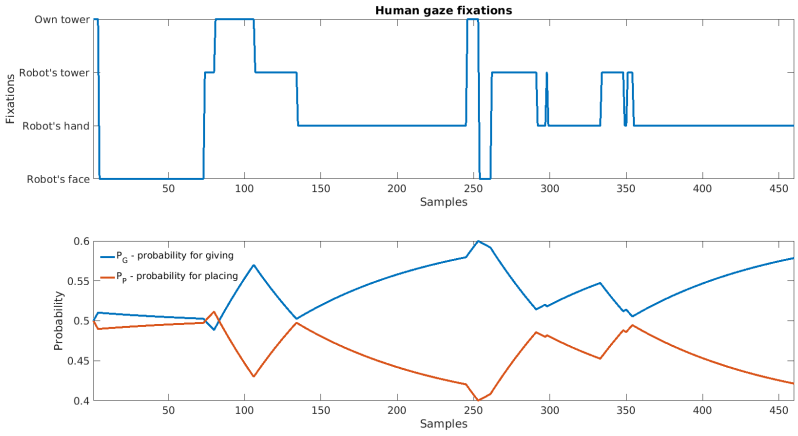

Fig. 7. Top: Human gaze fixations during the first HRI experiment. Bottom: The prediction of the understood action, i.e. the robot's understanding of the human behavior based on his gaze cues.

that time, the robot is continuing the non-verbal communication described in Section 4. This is an attempt of reaching action alignment with the human through the robot's gaze behavior. Since the robot does get any information from the human, i.e. no important cue provided by the eye tracker, the robot assumes the human did not yet understand the interaction intention, and will not complete the *giving* action. After the robot manages to catch the attention of the human, i.e. the human is looking at important cues of the interaction - states S_2 of the gaze behavior - the robot realizes the human understood the interaction intent, and proceeds to complete the handover action, see Fig. 7.

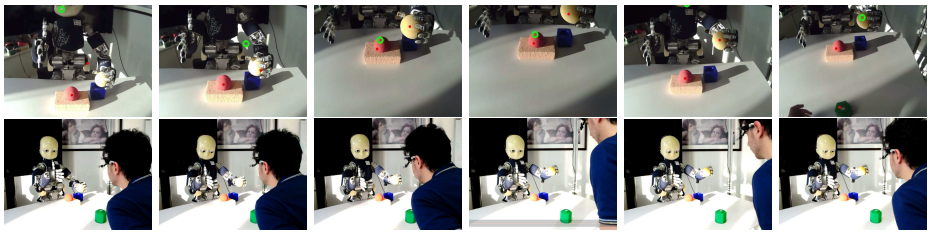

Fig. 8. The second experiment of a robot interacting with a human that misunderstands the robot's action. The interaction starts with an engaged human with the correct action, then the human misunderstands the robot's action, i.e. the action alignment, and hence, mutual alignment is broken. Only after looking at the robot, the human finally understands the actual robot action.

In the second experiment, we test the alignment of the robot, when the human misunderstands the action. Fig. 8 shows the human initially looking at the robot's face and hand. This implies that the human understands the on-going action, as it is seen from the action prediction outcome in Fig. 9.

The human then switches to fixate the robot's tower, see human gaze fixations in the top plot of Fig. 9 (samples [190-310]). This changes the prediction of the robot, concerning what the human understands, to a *placing* action. This results in the robot retracting the arm, signaling that there is no action alignment, and that the interaction needs to adapt. The human then looks again at the robot's face and hand, giving the robot the correct prediction of the action. The robot resumes the interaction and finally hands over the object. Supplementary video material is included for both interaction scenarios.

6 Conclusion and Future Work

This work describes a model of the stochastic gaze behavior of a leader, in a leader-follower social interaction. The gaze fixations are used as an instrument for non-verbal communication, to achieve transparency of the intended actions of an artificial agent. Simultaneously, the agent also reads the human partner's gaze cues to understand the action (s)he performs. Based on this feedback, an agent

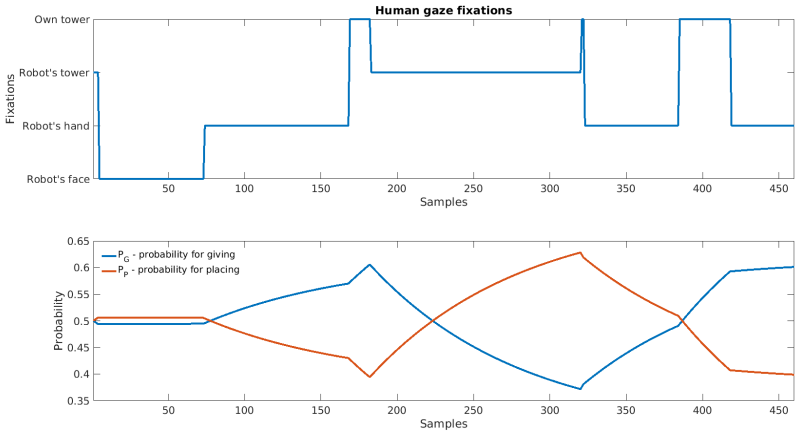

Fig. 9. Top: Human gaze fixations for the second HRI experiment. Bottom: Robot predictions of the human actions, updated over time. The robot adapts the arm movement in response to the human gaze behavior (Fig. 8).

can plan its motion to align its behavior to the current conditions of the social interaction. The proposed models for gaze behavior and action understanding, were integrated in the iCub's robot controller and validated in a HRI scenario with a human in the loop.

The iCub's gaze behavior was modeled with two discrete-time Markov chains, to drive the gaze before and after handover. The outcome of the models correlates to the analysis obtained from the HHI experiment data.

Inferring the level of understanding of the action by a human is also based on the HHI experiment data. From these data, an instantaneous probability of the two types of action (*giving* and *placing*) is built. These instantaneous probabilities integrated over time, are used to decide if the human understands the robot's action. Our experiments illustrate how the understanding of the action changes from the correct to the wrong action, and back again to the correct one. When the inferred action is misunderstood, it signals the robot to stop moving the arm toward the handover location, and to go back to the resting position. During that period, the gaze behavior continued to emit cues to communicate the intention of the interaction.

The future work will involve more thorough evaluation of the impact of the gaze behavior controller and motion planning alignment in the quality of HRI. We aim to enroll a group of naive subjects in a HRI with the iCub running the gaze behavior model and compared it to an alternative controller. It will allow us to analyze how the human gaze reaction time correlates with the understanding of the robot's action, and the initiation of the arm movement towards the handover location to take the object from the robot.

References

1. Admoni, H., Dragan, A., Srinivasa, S.S., Scassellati, B.: Deliberate delays during robot-to-human handovers improve compliance with gaze communication. In: Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction. pp. 49–56. HRI '14, ACM, New York, NY, USA (2014). https://doi.org/10.1145/2559636.2559682, http://doi.acm.org/10.1145/2559636.2559682
2. Andrist, S., Gleicher, M., Mutlu, B.: Looking coordinated: Bidirectional gaze mechanisms for collaborative interaction with virtual characters. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. pp. 2571–2582. CHI '17, ACM, New York, NY, USA (2017). https://doi.org/10.1145/3025453.3026033, http://doi.acm.org/10.1145/3025453.3026033
3. Bassetti, C.: Chapter 2 - social interaction in temporary gatherings: A sociological taxonomy of groups and crowds for computer vision practitioners. In: Murino, V., Cristani, M., Shah, S., Savarese, S. (eds.) Group and Crowd Behavior for Computer Vision, pp. 15 – 28. Academic Press (2017). https://doi.org/https://doi.org/10.1016/B978-0-12-809276-7.00003-5, http://www.sciencedirect.com/science/article/pii/B9780128092767000035
4. Biagini, F., Campanino, M.: Discrete time markov chains. In: Elements of Probability and Statistics, pp. 81–87. Springer (2016)
5. Domhof, J., Chandarr, A., Rudinac, M., Jonker, P.: Multimodal joint visual attention model for natural human-robot interaction in domestic environments. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 2406–2412 (Sept 2015). https://doi.org/10.1109/IROS.2015.7353703
6. Duarte, N.F., Rakovi, M., Tasevski, J., Coco, M.I., Billard, A., Santos-Victor, J.: Action anticipation: Reading the intentions of humans and robots. IEEE Robotics and Automation Letters **3**(4), 4132–4139 (Oct 2018). https://doi.org/10.1109/LRA.2018.2861569
7. Duchowski, A.T.: Gaze-based interaction: A 30 year retrospective. vol. 73, pp. 59 – 69 (2018). https://doi.org/https://doi.org/10.1016/j.cag.2018.04.002, http://www.sciencedirect.com/science/article/pii/S0097849318300487
8. Farha, Y.A., Richard, A., Gall, J.: When will you do what? - anticipating temporal occurrences of activities. arXiv preprint arXiv:1804.00892 (2018)
9. Fathi, A., Ren, X., Rehg, J.M.: Learning to recognize objects in egocentric activities. In: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3281–3288. CVPR '11, IEEE Computer Society, Washington, DC, USA (2011). https://doi.org/10.1109/CVPR.2011.5995444, http://dx.doi.org/10.1109/CVPR.2011.5995444
10. Gallotti, M., Fairhurst, M., Frith, C.: Alignment in social interactions. Consciousness and cognition **48**, 253–261 (2017)
11. Gottwald, J.M., Elsner, B., Pollatos, O.: Good is upspatial metaphors in action observation. Frontiers in Psychology **6**, 1605 (2015). https://doi.org/10.3389/fpsyg.2015.01605, https://www.frontiersin.org/article/10.3389/fpsyg.2015.01605
12. Ivaldi, S., Anzalone, S., Rousseau, W., Sigaud, O., Chetouani, M.: Robot initiative in a team learning task increases the rhythm of interaction but not the perceived engagement **8**, 5 (02 2014)

13. Kassner, M., Patera, W., Bulling, A.: Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In: Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing: Adjunct publication. pp. 1151–1160. ACM (2014)
14. Kelley, R., Tavakkoli, A., King, C., Nicolescu, M., Nicolescu, M.: Understanding activities and intentions for human-robot interaction (2010). https://doi.org/10.5772/8127, https://doi.org/10.5772/8127
15. Kitani, K.M., Ziebart, B.D., Bagnell, J.A., Hebert, M.: Activity forecasting. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) Computer Vision – ECCV 2012. pp. 201–214. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)
16. Koppula, H.S., Saxena, A.: Anticipating human activities using object affordances for reactive robotic response. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(1), 14–29 (Jan 2016). https://doi.org/10.1109/TPAMI.2015.2430335
17. Kothe, C.: Lab streaming layer (lsl). https://github. com/sccn/labstreaminglayer. Accessed on February **26**, 2015 (2018)
18. Lukic, L., Santos-Victor, J., Billard, A.: Learning robotic eye–arm–hand coordination from human demonstration: a coupled dynamical systems approach. Biological cybernetics **108**(2), 223–248 (2014)
19. Metta, G., Natale, L., Nori, F., Sandini, G., Vernon, D., Fadiga, L., Von Hofsten, C., Rosander, K., Lopes, M., Santos-Victor, J., et al.: The icub humanoid robot: An open-systems platform for research in cognitive development. Neural Networks **23**(8-9), 1125–1134 (2010)
20. Palinko, O., Rea, F., Sandini, G., Sciutti, A.: Eye gaze tracking for a humanoid robot. In: 2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids). pp. 318–324 (Nov 2015). https://doi.org/10.1109/HUMANOIDS.2015.7363561
21. Pattacini, U., Nori, F., Natale, L., Metta, G., Sandini, G.: An experimental evaluation of a novel minimum-jerk cartesian controller for humanoid robots. In: Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on. pp. 1668–1674. IEEE (2010)
22. Pfeiffer, M., Schwesinger, U., Sommer, H., Galceran, E., Siegwart, R.: Predicting actions to act predictably: Cooperative partial motion planning with maximum entropy models. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 2096–2101 (Oct 2016). https://doi.org/10.1109/IROS.2016.7759329
23. Rakovic, M., Duarte, N., Marques, J., Santos-Victor, J.: Modelling the gaze dialogue: Non-verbal communication in human-human and human-robot interaction. Paper under revision **1**(1), 1–12 (2018)
24. Raković, M., Duarte, N., Tasevski, J., Santos-Victor, J., Borovac, B.: A dataset of head and eye gaze during dyadic interaction task for modeling robot gaze behavior. In: MATEC Web of Conferences. vol. 161, p. 03002. EDP Sciences (2018)
25. Roncone, A., Pattacini, U., Metta, G., Natale, L.: A cartesian 6-dof gaze controller for humanoid robots. In: Robotics: Science and Systems (2016)
26. Schydlo, P., Rakovic, M., Jamone, L., Santos-Victor, J.: Anticipation in human-robot cooperation: A recurrent neural network approach for multiple action sequences prediction. Robotics and Automation. ICRA 2018. IEEE International Conference on Robotics and Automation (2018)

27. Sciutti, A., Mara, M., Tagliasco, V., Sandini, G.: Humanizing human-robot interaction: On the importance of mutual understanding. IEEE Technology and Society Magazine **37**(1), 22–29 (March 2018). https://doi.org/10.1109/MTS.2018.2795095
28. Ycel, Z., Salah, A.A., Merili, ., Merili, T., Valenti, R., Gevers, T.: Joint attention by gaze interpolation and saliency. IEEE Transactions on Cybernetics **43**(3), 829–842 (June 2013). https://doi.org/10.1109/TSMCB.2012.2216979
29. Zhang, J., Li, W., Ogunbona, P.O., Wang, P., Tang, C.: Rgb-d-based action recognition datasets: A survey. Pattern Recognition **60**, 86 – 105 (2016). https://doi.org/https://doi.org/10.1016/j.patcog.2016.05.019, http://www.sciencedirect.com/science/article/pii/S0031320316301029