Action Anticipation: Reading the Intentions of Humans and Robots

Nuno Ferreira Duarte, Mirko Raković, Jovica Tasevski, Moreno Ignazio Coco, Aude Billard, and José Santos-Victor

***Abstract*—Humans have the fascinating capacity of processing nonverbal visual cues to understand and anticipate the actions of other humans. This "intention reading" ability is underpinned by shared motor repertoires and action models, which we use to interpret the intentions of others as if they were our own. We investigate how different cues contribute to the legibility of human actions during interpersonal interactions. Our first contribution is a publicly available dataset with recordings of human body motion and eye gaze, acquired in an experimental scenario with an actor interacting with three subjects. From these data, we conducted a human study to analyze the importance of different nonverbal cues for action perception. As our second contribution, we used motion/gaze recordings to build a computational model describing the interaction between two persons. As a third contribution, we embedded this model in the controller of an iCub humanoid robot and conducted a second human study, in the same scenario with the robot as an actor, to validate the model's "intention reading" capability. Our results show that it is possible to model (nonverbal) signals exchanged by humans during interaction, and how to incorporate such a mechanism in robotic systems with the twin goal of being able to "read" human action intentionsand acting in a way that is legible by humans.**

***Index Terms*—Social human-robot interaction, humanoid robots, sensor fusion.**

Fig. 1. Human-Human Interaction: an experiment involving one actor (top-right) *giving* and *placing* objects and three subjects reading the intentions of the actor (left); Human-Robot Interaction: a robot performing the human-like action and subjects try to anticipate the robots' intention (bottom-right).

I. Introduction

WHEN working in a shared space, humans interpret nonverbal cues such as eye gaze and body movements to understand the actions of their workmates. By inferring the actions of others, we can efficiently adapt our movements and appropriately coordinate the interaction (Fig. 1). According to Dragan *et al.* [1], the intention of others can only be understood if and when the end-goal location becomes unambiguous to us. For that same reason, to improve human-robot interaction (HRI), robots should perform coordinated movements of all body parts, so that their actions and goals can be "legible" to humans.

Recent research in HRI has focused on studying the human behaviour [2]–[5]. Several papers, which we will discuss in more detail in Section II, have built bio-inspired controllers that facilitate human action understanding and interaction, and improve the communication with robots. However, they do not focus on the essential part of human interaction - the communication of intent - the central focus of our work.

We start by defining a scenario of human-human interaction (HHI), detailed in Section III, to study non-verbal communication cues between humans, in a quantitative manner. The experiment consists of an actor performing goal-oriented actions in front of three humans sitting at a round table (Fig. 1-left). The actor picks up a ball placed in front of him and has to

Manuscript received February 23, 2018; accepted July 11, 2018. Date of publication July 31, 2018; date of current version August 17, 2018. This letter was recommended for publication by Associate Editor P. Falco and Editor D. Lee upon evaluation of the reviewers' comments. This work was supported by EU H2020 project under Grant 752611—ACTICIPATE, in part by the FCT project UID/EEA/50009/2013, and in part by the RBCog-Lab research infrastructure. *(Corresponding author: Nuno Ferreira Duarte.)*

N. Duarte, M. Raković, and J. Santos-Victor are with the Vislab, Institute for Systems and Robotics, Instituto Superior Técnico, Universidade de Lisboa, Lisbon 1649-004, Portugal, and also with the Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Novi Sad 21000, Serbia (e-mail: nferreira duarte@isr.tecnico.ulisboa.pt; rakovicm@isr.tecnico.ulisboa.pt; jasv@isr.tecnico.ulisboa.pt).

J. Tasevski is with the Faculty of Technical Sciences, University of Novi Sad, Novi Sad 21000, Serbia (e-mail: tasevski@uns.ac.rs).

M. Coco is with the Department of Psychology (Centre for Cognitive Ageing and Cognitive Epidemiology), University of Edinburgh, Edinburgh EH8 9JZ U.K. (e-mail: moreno.coco@ed.ac.uk).

A. Billard is with the Learning Algorithms and Systems Laboratory, School of Engineering, École Polytechnique Fédérale de Lausanne, Lausanne 1015, Switzerland (e-mail: aude.billard@epfl.ch).

This letter has supplemental downloadable multimedia material available at http://ieeexplore.ieee.org, provided by the authors. The Supplementary Materials contain a video illustration of the RAL work. This material is 8.49 MB in size.

Digital Object Identifier 10.1109/LRA.2018.2861569

either (i) *place* the ball on the table in front of one of the three persons or (ii) *give* the ball to one of them (Fig. 1-top right). Considering the two actions (placing/giving) and three spatial parametrizations (left/middle/right), the actor executes one out of six action-possibilities. With this HHI experiment, we have built a dataset with the actor's 3D body movements and eye-gaze information during the interaction. Additionally, video recordings were taken during the entire experiment, and used to design a human study.

The videos of the actor are used to analyse three different cues: eye gaze, head orientation, and arm movement towards the goal position (Section IV) of *placing* and *giving* actions. For this study, we prepared a gated experiment, using a set of video segments of increasing temporal duration, of each action performed by the actor. The video fractions are shown to the participants, and they are asked to predict the actor's intended action: *giving* the ball to one of the persons or *placing* the ball at one of three assigned markers on the table (6 possibilities in total). Our results reveal that early eye-gaze shifts provide important information for the human subjects to anticipate the intention of the actor. Additionally, we observed significant of the eye-gaze behaviour between *giving* and *placing* actions, that seems to be governed by and attend to multiple goals.

The recordings of the upper body and eye gaze motion are used to develop a computational model of the human actions (Section V). The arm movement was modelled with Gaussian Mixture Models (GMM), and Gaussian Mixture Regression (GMR) is used to generate the arm trajectory. The eye gaze behaviour depends on the type of action. Before picking up the ball, the eye fixates the initial ball position. Then, for the *placing* action, the eye gaze aims at the goal position (i.e. marker on the table). In case of the *giving* action, the eye gaze switches between the face of the human and end-goal position (i.e. the handover location).

The developed computational model is incorporated in a controller for the iCub humanoid robot, with the purpose of validating the model and investigating whether humans can "read" the robot actions in the same way they can read the actions of other humans. We have built a second human study for the same scenario using a robot actor. We recorded videos of the robot performing the same set of actions as the human actor. The video fractions of the robot-actor are then presented to another group of participants, who are asked to anticipate the robot's action intention (Section VI).

In Section VII we discuss our experiments and results concerning the human perception of the robot's actions, in terms of readability. Our results show that we can model the non-verbal communication cues during human-human interaction and transfer that model to a robot executing *placing* actions or *giving* a ball to a human. Finally, we draw some conclusions and establish directions of future work.

II. State of the Art

Dragan *et al.* [1] discuss the aspects of predictability and legibility of arm movements. They define legible robot actions as copies of human actions but executed with exaggerated movements, and demonstrate that they can be understood sooner. Instead, in our work, legibility is not achieved by exaggerating the arm movements, but by modelling the natural coordination of human eye, head, and arm movements. For that purpose, we conduct a quantitative analysis of the importance of the robot eye-gaze behaviour for the legibility of the robot's movements. We validate the model with a human study where subjects need to read the robot's intentions and select between (*placing*) or (*giving*) actions with three spatial parametrizations.

Research in HRI and, more specifically, in human motion understanding [6]–[8] and modelling [9], has relied on different existing datasets. Zhang *et al.* [9] present a survey on RGB-D based action recognition dataset. The CAD 120 dataset [10] includes a rich repertoire of human actions including the labels of the activities performed during those actions. Some of the existing datasets only provide information related to 3D body coordinates, while the few which include gaze tracking have the drawback of being limited to 1 or 2 tasks [11]–[13]. The first contribution of this letter is to provide a publicly-available dataset, that overcomes the shortcomings of existing datasets and contains synchronised and labelled video+gaze and body motion in a dyadic scenario of interaction.[1] This dataset has already been successfully used to develop a novel action anticipation algorithm, that integrates the cues from both gaze and body motion to provide faster and more accurate predictions of human's action [14].

Neurobiology provides extensive insight into the biological models of the human sensory-motor system. One group of neuroscientists have focused on investigating cortical structures such as the posterior parietal cortex, the premotor and the motor cortices [15]. Another stream of research has been directed on modelling the role of the cerebellum in the motor loop, movement generation and synchronisation of sensory-motor system [16]. These findings are used in [17], [18] to develop coupled dynamical systems framework for arm-hand and eye-arm-hand motion control for robots. The framework is focused on motor control coupling. Here, we extend our previous work, to the analysis of the interpersonal coordination of sensory-motor systems during interaction. Therefore our dataset of coordinated gaze and body movements during dyadic interactions is then used to build a bio-inspired model.

Authors in [19] investigate the infants' perception during object-handover interactions. Those studies show that, in spite of their young age, the gaze behaviour is already modulated by the social interaction context. The work described in [20] shows how the gaze behaviour encompasses multiple fixation points when the subject is engaged in complex tasks, such as tea-making. However, none of these works develops experiments with on-line tracking of the eye gaze, head orientation, and arm movements during an interpersonal interaction, with *placing* or *giving* actions in different spatial parametrizations.

Meng *et al.* [21] study human eye-gaze during interaction. They built an experiment where different types of gaze trajectories are examined in a human-robot scenario. However, their

[1]The dataset of synchronised video, gaze fixations from Pupil eye tracker, and body motion from OptiTrack motion tracking system of *placing* and *giving* actions can be downloaded from: http://vislab.isr.tecnico.ulisboa.pt/datasets/#acticipate1.

analysis is not based on a quantitative sensory system but, rather, by manually labelling at the subject's eyes in the video recordings. We propose using an eye-tracking system to record and assess the human gaze behaviour in those actions. Furthermore, they conclude that, for *giving* actions, humans prefer when the robot fixates the person's face and then switches, i.e. looks, to the handover position, as opposed to just looking either the face or the handover position exclusively. This is a contextually based behaviour of the gaze that we intend to study using the eye-tracking system.

The second set of limitations in [21]–[23] concerns the robot used in the experiments. Due to the limited number of degrees of freedom in the head of the robot, the eye gaze shifts are simulated with head rotation. In our work, we use an eye-tracking device to observe the actual gaze fixation points during the interaction independently from the head gaze as this provides better accuracy than just the head orientation [24]. We use the iCub humanoid robot that has a human-like face where the eyes can independently move, and thus express a readable behaviour of eye-gaze and head-gaze.

III. Interaction Scenario

This section presents an interaction scenario for collecting: (i) videos of actor movements to study the contribution of different cues and timings on anticipation of actions and (ii) the motion of the eye-gaze and relevant body-parts of a human actor, to model the human movements.

A. Scenario Description

The scenario can be seen in Fig. 1(left). For each trial, one actor executes a set of *placing* or *giving* actions directed towards one of the three (left/middle/right) subjects. The actor was instructed to act as normal as possible when performing those actions. The actor picks the object from the initial position and executes one of these 6 preselected action-configurations (2 actions and 3 spatial directions).

- *placing* on the table to the actor's **left** (P_L), **middle** (P_M), or **right** (P_R),
- *giving* the ball to the person on actor's **left** (G_L), **middle** (G_M), or **right** (G_R).

The actions to execute were instructed over an earpiece to the actor so that none of the other participants could know which would be performed next. The order of the actions is randomly selected to prevent the actor from adapting its posture prior to initiation. Every action begins with picking up the ball and ends with the actor placing the ball back to the initial position on the table.

B. Hardware and Software Setup

The actor movements were recorded with an OptiTrack motion capture (MoCap) system, consisting of 12 cameras all around the environment and a suit with 25 markers, placed on the upper torso, arms, and head, that is worn by the actor. The MoCap provides position and orientation data of all relevant body parts (head, torso, right-arm, left-arm).

The eye gaze was recorded with the mobile, binocular Pupil-Labs eye tracker [25], that allowed us to track the actor's fixation point. To track the head movements with the MoCap system, head markers were placed on the Pupil-Lab system. To record the scene, three video cameras are used to provide different viewing angles that will complement during the evaluation phase. The first camera provides the world-view perspective of the actor from the Pupil Labs eye tracking headset (top-right image in Fig. 1, the small window on top). The second camera records the table top where the actions will take place. This one provides a continuous look at the table and all the actor's movements (Fig. 1-top right). The third camera was located further from the scene, looking inwards, giving a proper reading of the subject's actions and an outlook of the experiment (Fig. 1-left).

To collect all the sensory information, the OptiTrack's Motive and Pupil Lab's Pupil Capture software were used. Prior to recording, both sensors were calibrated. Custom software was developed to acquire the video of the actor's action. All the sensory data are captured on distributed machines and data are streamed through the Lab Streaming Layer [26] for centralised storage and data synchronisation.

C. Synchronization of Sensory Data

A total of 120 trials are performed with action-configurations: P_L, P_M, P_R, G_L, G_M and G_R performed 20, 23, 17, 17, 19 and 24 times respectively. The binocular eye gaze tracking system recorded world camera video and eye gaze data at 60 Hz, the motion capture system recorded the movements of the body at 120 Hz, and video camera facing the actor, recorded video at 30 Hz. The data from all sensing systems are streamed and collected at one place, with the timestamps of each sensing system as well as the internal clock information, that is used as a reference to synchronise all sensory flows.

IV. Reading the Intentions of Humans

We conducted a human study to quantify how the different cues contribute to the ability to anticipate the actions of others, and how those cues are related to the spatial (left/middle/right) distribution. The study includes a questionnaire pertaining to the actions performed by an actor.

A. Participants

The study involved 55 participants (40 male, and 15 female), age 31.9 $\pm$ 13 (mean $\pm$ SD). There were 13 teenagers and 6 people over 50 years of age. Approximately 62% were students, 27% were professors, 7% were researchers, and 4% were staff members, 3 subjects were left-handed. All subjects were naive with respect to the purpose of the research.

B. Human Study

The subjects were presented with videos of an actor performing *giving* or *placing* actions in the different spatial directions, and were asked to reply to a questionnaire related to the action being executed. The questionnaire consists of 24 questions.[2]

[2]A description of the human study can be seen at the following web address: http://vislab.isr.tecnico.ulisboa.pt/wp-content/uploads/2018/07/acticipate1_questionnaire_description.pdf

Before the question is shown to the subject, they have to watch a short video of the actor performing one of the six possible actions (2 end-goal actions multiplied by 3 directional end-goal locations). Based on the video shown, the participant had to identify the actor's intended action. The videos were fractioned into four types according to the cues provided by the eye gaze shift, head gaze shift, and arm movement. This can be understood as a gated experiment in which fractions of video segments are shown to subjects beginning when the actor grabs the object and ending when:

- there is a saccadic eye movement towards the goal - G
- "G" plus the head rotates to the same goal - G+H
- "G+H" plus the arm starts moving to the goal - G+H+A
- "G+H+A" plus the arm finishes the trajectory to the goal - G+H+A+.

The last group of videos (G+H+A+) was used as a golden standard to remove outliers. Out of 24 questions, the first three were used to familiarise participants with the questionnaire and were discarded from the analysis. Out of the remaining 21 questions, five questions are from the G difficulty level, six are from G+H, six are from G+H+A, while four were used for detecting outliers. Twelve are for *placing* and nine are for *giving* actions, whereas seven belong to left, eighth to middle and six to right direction.

C. Analysis

From Section IV-B we reach 5 important conclusions, that we describe in the following paragraphs.

The first conclusion is the most obvious and is shown in Fig. 2(a). The more temporal information is available to subjects, the better the decision is, the higher the success rate and the lower the variance. We validate this trend with a quantitative analysis, with a two-way ANOVA [27], that shows a very significant correlation between the amount of information and the success rate, $F(2\,5560) = 1396.76$, $p < 0.0001$. Gaze alone is responsible for a 50% success rate of (about 3 times the chance level of 1/6 = 16.7%).

The analysis is further refined by considering two variations: (i) how well can the subjects predict spatial orientation, irrespective of the *giving* vs *placing* action? and (ii) how can the subjects predict the action (*giving*, or *placing*) irrespective of the orientation (**left**, **middle**, or **right**)?

Secondly, according to our results, the prediction of spatial orientation does not depend strongly on the amount of temporal information. The participants did not report significant difficulties to understand the gaze orientation from the "G" videos when the actor was wearing the eye tracker, compared to a case where no glasses are used. Gaze alone is crucial for action understanding in the azimuth orientation, 85% success (chance level of 33%), then head information only increases around 15%. Instead, action prediction depends strongly on the amount of temporal information. Surprisingly, subjects were only capable of understanding the action-type 60% (chance level of 50%) of the time for the first video fraction, but as more information was provided the success rate increased quite rapidly. To analyse in more detail the reason why, we refined this results in Fig. 2(b) to study two conditions: (i) *giving* actions and (ii) *placing* action.

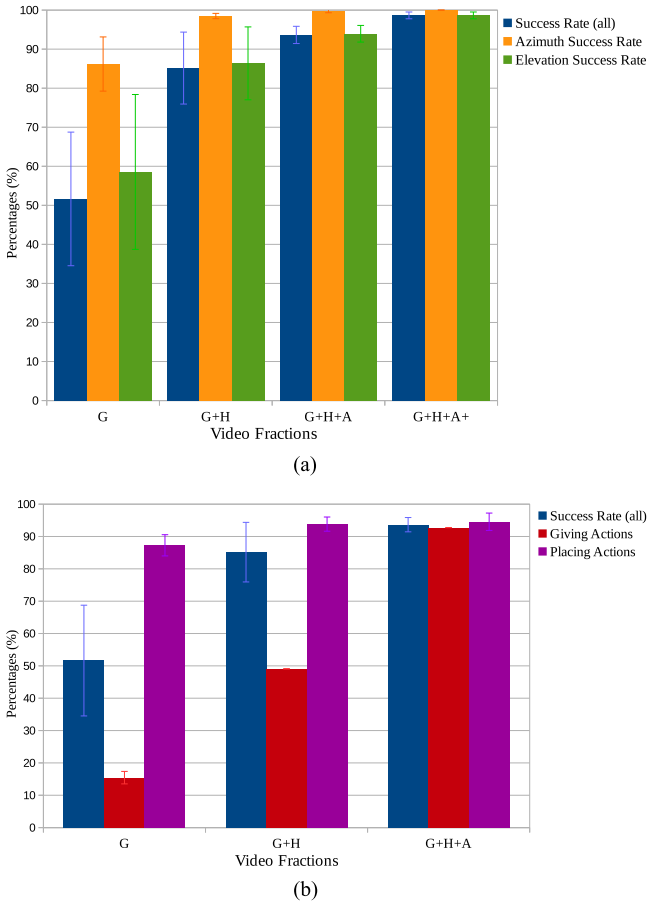

Fig. 2. The success of the participants identifying the correct action a) overall success rate; success rate in identifying the direction of the action; b) success rate in identifying the *giving* and *placing* actions.

Thirdly, we observe a significant interaction between the type of action and the amount of information available to the subject. This is confirmed by the two-way ANOVA, $F(2\,5560) = 537.70$, $p < 0.0001$. For the *placing* action we have a success rate of 85% (chance level 50%) with gaze alone. However, we observe that for the *giving* action we get a success rate lower than chance level. Our fourth conclusion comes from the two-way ANOVA, confirming a significant importance between type of action and subjects' success rate, $F(1\,5560) = 2306.78$, $p < 0.0001$, indicating a bias towards *placing* in this HHI scenario.

These experiments clearly demonstrate, quantitatively, the importance of gaze in a dyadic action. In HHI, eye gaze information provides the necessary information to predict the intention of the other subject. For *giving* actions this is not the case, but we believe that the experimental setup geometry introduces a unintentional bias towards the action that requires the least energy, *placing* the object on the table. We evaluate the bias towards *placing* by showing additional videos segmented before any non-verbal cue (smaller than "G" video fraction) and the results show that in the case of *placing* vs *giving*, the majority of people picked *placing*, proving a significant preconception in this HHI scenario. Our final conclusion is our cornerstone of this letter. This analysis shows that human eye-gaze provides key information to read the action correctly, and justifies the

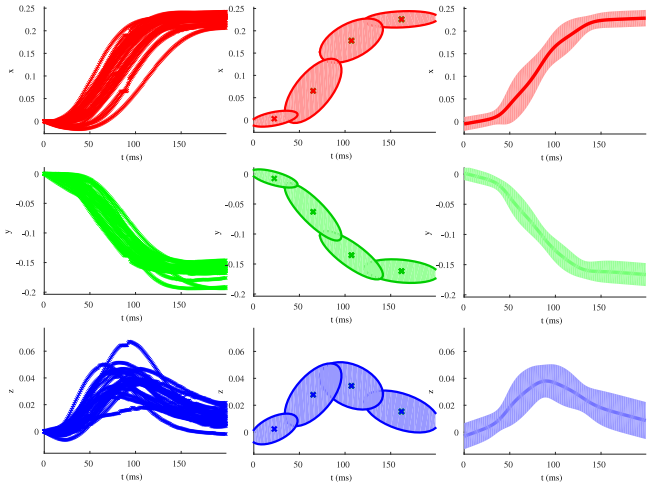

Fig. 3. Recorded coordinates of human hand performing P_R action, representation of corresponding covariance matrices and output from GMR with covariance information.

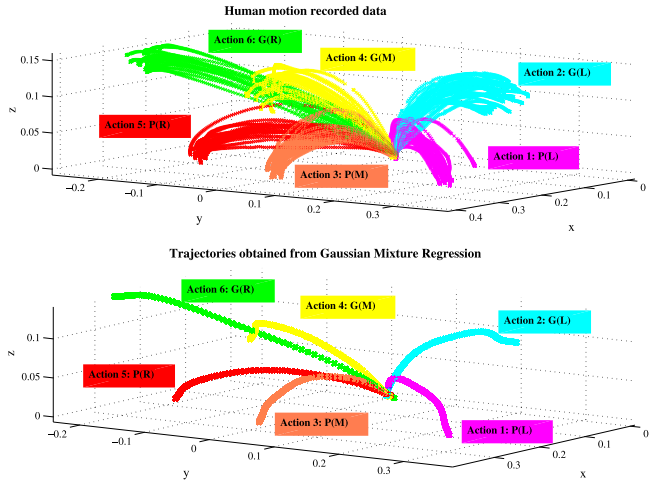

Fig. 4. Spatial distribution of hand motion for all six actions (top) and corresponding output from GMR (bottom).

need to include human-like, eye-gaze control, in order to improve action-legibility and anticipation as required for efficient human-robot interaction.

V. Modeling Human Motion

This section begins by explaining the modelling of the arm motion and then proceeds by analysing the eye movements.

We use a Gaussian Mixture Model (GMM) [28] to model the trajectories of the arm movement in a probabilistic framework. The motion is represented as a state variable $\{\xi_j\}_{j=1}^{N} \in \mathbb{R}^3$, where N is the total number of arm trajectories for all actions, and ξ_j are the Cartesian coordinates of the hand for *giving* or *placing* actions. The GMM defines a joint probability distribution function over the set of data from demonstrated trajectories as a mixture of k Gaussian distributions each one described by the prior probability, the mean value and the covariance matrix.

$$\begin{aligned} p(k) &= \pi_k \\ p(\xi_j|k) &= \mathcal{N}(\xi_j; \mu_k, \Sigma_k) \\ &= \frac{1}{\sqrt{(2\pi)^D |\Sigma_k|}} e^{-\frac{1}{2}((\xi_j-\mu_j)^T \Sigma_k^{-1}(\xi_j-\mu_j))} \end{aligned} \quad (1)$$

where $\{\pi_k; \mu_k, \Sigma_k\}$ is the prior probability, mean value, and covariance, respectively, for each k normal distribution.

The left column in Fig. 3 shows an example of the recorded trajectories of the actor's hand during execution of the P_R action. The middle column shows the recorded trajectories encoded in GMM, with covariances matrices represented by ellipses. We use four Gaussian distributions to model the behaviour of the arm trajectory for each Cartesian coordinate. This is to take into account the minimum error and the increase of complexity of the problem. Then the signal is reconstructed using Gaussian Mixture Regression (GMR). The new parameters, mean and covariance for each Cartesian coordinate, are defined as in [28]. The right column represents the GMR output of the signals in bold and the covariance information as the envelope around the bold line.

The same modelling is done for all the 2 actions and 3 orientations. Fig. 4-top, shows the spatial distribution of the recorded data for all six actions represented by six different colours. Fig. 4-bottom, shows the spatial distribution of modelled actions obtained with GMR.

Moon *et al.* [29] observed that the human eye gaze exhibits a switching behaviour during *giving* actions. This was observed in a HHI experiment scenario where two humans are *giving* a bottle to each other. The work has several shortcomings. First, the experiment can not guarantee that gaze behaviour occurs in general settings. Once the human knows which action will take place, there is no need to infer the action from non-verbal communication. Secondly, the analysis of the different gaze behaviours was done empirically (manually labelling the videos). In our dataset, we have measured fixation points, the actual points of interest in a handover task, and the duration of eye gaze between each switching behaviour. As future work, we use this information to design a detailed biologically-inspired, eye-gaze controller for HRI scenarios.

In the HRI experiment of [29], when the robot gaze fixations switches from the human's face to the handover position, it does not improve the speed of the human reaching time, but it does improve the perception of the interaction. This corroborates the findings in [21]. Our work studies the same behaviour, using a humanoid robot and eye-gaze cues extracted from the HHI experiment.

The information collected to model the human gaze behaviour, was acquired with an eye-tracking system (Pupil-Labs). Fig. 5 shows five different cases of the spatio-temporal distribution of the fixation point marked with a green circle. Fig. 5(a) shows the spatio-temporal distribution of fixation points for the P_M *placing* action in which the green circle is concentrated around the goal position of the red ball.

Fig. 5(b)–(e) show the spatio-temporal distributions of the fixation points during G_M *giving* action when the actor was fixating: (i) only the hand of the person, (ii) only the face of the person, (iii) first the hand and then the face, and (iv) first the face and then the hand. From this observed behaviour, we

Fig. 5. Sequence of images of spatiotemporal distribution of fixation point for *placing* and *giving* actions. Subgroup (a) is related to action P_M. The actor only fixates the center marker which is the end-goal point for the action. Subgroups (b)–(e) correspond to action G_M. The actor changes fixation point in 4 different patterns: (b) actor's only fixates the hand of the subject in front; (c) only fixating the subject in front; (d) it begins by fixating the subject's hand and it ends by fixating the subject's eyes; (e) it fixates the subject's eyes in the beginning and it ends the fixation by looking at the subject's hand.

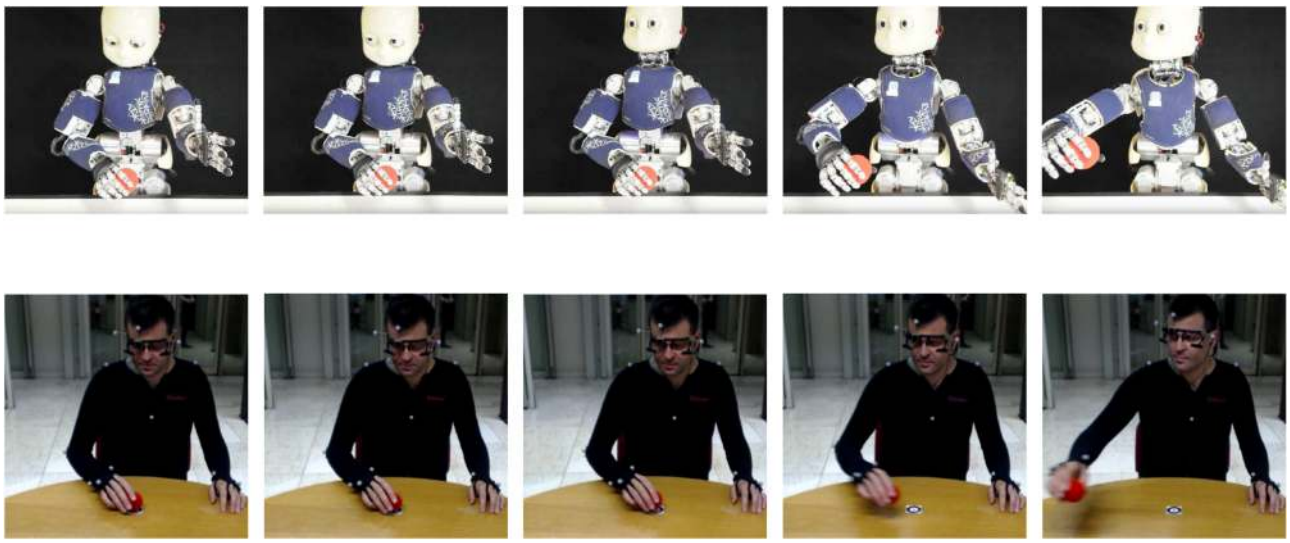

Fig. 6. The sequence of images of a robot (top) and an actor (bottom) performing the G_R action. The first sequence is the initial point for both the actor and the robot. The second stage corresponds to when the short video stops at the video fraction 'G'. The third is at video fraction "G+H." Forth and fifth sequences are for the final two video fractions, corresponding to the arm motion.

designed a controller that will generate an equivalent switching behaviour of the fixation point, i.e. a qualitatively similar eye-gaze behaviour.

The robot gaze controller was implemented as a state-machine that (qualitatively) replicates the gaze shift behaviour observed during human-human interaction. The controller's initial state is the starting location of the ball. Then, depending on the action, there is a state transition to the final location of the ball (*placing*) or a switch between two states: (i) face of the person, (ii) handover location, (*giving*). The desired fixation point is input to the coupled eye-head controller that executes saccadic eye movements, followed by the coordinated motion of the eye/neck joints. Fig. 6 shows the sequence of images, during the execution of the G_R action by the iCub robot and the corresponding images of the actor, when the actor looks first to hand of the other person and then switches to the face.

The validation of the controller is presented in Section VI. The reference arm trajectory is generated with a GMR and the arm's joints are controlled with a minimum jerk Cartesian controller. The robot eye controller was based on the qualitative analysis of the human gaze behaviour and the eye's and neck joints are simultaneously controlled using Cartesian 6-DOF gaze controller [30].

VI. Reading the Intentions of Robots

To study the readability of robot's intention, we prepared a second questionnaire with the same set of actions performed by a robot. To assess the relative importance of the different non-verbal (eye, head, arm) cues we have added new conditions: (i) blurring the eyes in the video, and (ii) blurring the entire head.

This second human study involved 20 participants answering 36 questions: 18 without any blurring; 12 with eye blurring, and 6 with the whole head blurred. The 12 eyes blurred questions correspond to the 6×2 possible action-configurations with: (i) blurred eye gaze and with visible head gaze and (ii) blurred eye gaze, with visible head gaze, and visible arm movement); the 6 whole head blurred questions correspond to the 6×1 possible action-configurations with blurred eye gaze, blurred head gaze, and visible arm movement). There were less participants in this second study but each subject had to answer to more questions than in the previous case. Fig. 7(a) shows the participants success rate in identifying the robot-action in the three cases: *giving* action, *placing* action or both.

As in the first study, we observed that the more temporal information the subjects had, the better their decision was, and the higher the success rate. The average success rate increases as more information is provided, Fig. 7(a).

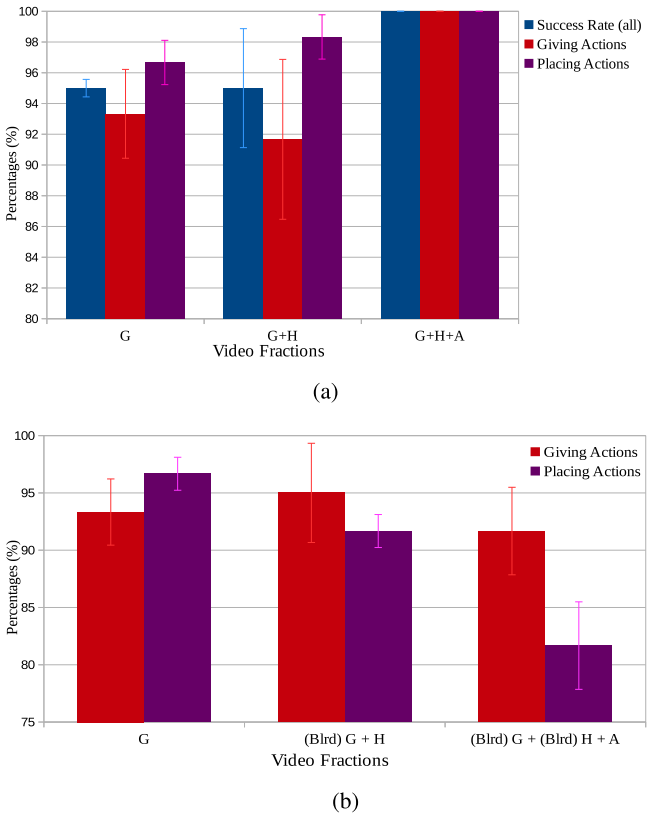

Fig. 7. The success of the participants identifying the correct action: (a) Comparison between the overall success, *giving* actions, and *placing* actions, throughout all the different video fractions; (b) The effects of blurring in the success rate. The "(Blrd)" indication after one non-verbal cues, means that one or more non-verbal cues were blurred, in those video fractions. For example, "(Blrd) G + H" means that the eyes were blurred in those video fractions, thus preventing the use of gaze information to read the robot's intention. Nevertheless, in those videos the orientation of the head was still visible. The y axis starts from 80% for visualization purposes. Gaze and head information (G + H) is in (a), success rate for *giving* or *placing* actions for the blurred gaze and just head information (blurred G + H) is in (b)

In addition, we analyse the effects of blurring on the success rate of *placing* and *giving* actions, Fig. 7(b). We can see that when blurring the eyes, and preserving only the head information ("(Blrd) G+H") the success rate drops around 5%. Since there is a clear distinction between the head orientation in *placing* and *giving* actions, for most people, this is enough information to predict the robot's intention. When blurring the whole head, the only information available is the motion of the arm. In [1], this motion is classified as "predictable," as such, it will not give the most information to the user.

Our experiment showed that the difficulty increases with the increase of the blurred area. This means that the legibility of the robot's actions improves with the integration of human-like eye gaze behaviour into the controller. Our work generalizes Dragan *et al.* [1], as legibility is achieved through the combination of both human arm, body, and eye-gaze movements.

Our results show the importance of non-verbal cues in a human-human interaction scenario, and we successfully transferred the models to a human-robot experiment, where human-level action-readability of robot actions was achieved.

VII. Discussions and Conclusion

We conducted experiments and studies to investigate the human ability to read the intention of a human actor during *placing* and *giving* actions in 6 pre-selected action-configurations. One of our contributions was a publicly available dataset of synchronized videos, gaze and body motion data. We conducted an HHI experiment with two objectives: (i) understand how the participants manage to predict the observed actions of the actor; (ii) use the collected data to model the human arm behaviour and (in a qualitative sense) the eye gaze behaviour.

With the human study data, we analysed the different types of non-verbal cues during interpersonal interaction: eye gaze, head gaze, and arm information. For the *placing* actions, with just eye gaze information 85% of subjects can read the intentions of the actor correctly (chance 50%). However, for the *giving* actions the results were much worse (chance 20%). To understand the reason behind these results we analysed the eye gaze behaviour recordings from the eye-tracking system for the *giving* actions. The analysis of these data shows that for the same type of action, there are different gaze trajectories Fig. 5(b)–(e). According to Moon *et al.* [29], humans prefer a *giving* action when the actor performs this switching behaviour [21] observed in Fig. 5(d)–(e). This switching behaviour can be seen as a confirmation routine to acknowledge to the other person that an interaction is taking place. Since the human motion is a combination of eyes, head, and arm movement, coupled during the action execution, the "communication" is only properly established, once it is signalled with this behaviour. As such, the logical choice is to infer that the actor is not trying to communicate with us, which justifies the preference for the *placing* action.

After the analysis, our next contribution was on modelling the human behaviour from the data collected. The arm movements of the actor were modelled with GMM/GMR that can replicate the natural movement of the human arm. Dragan *et al.* [1] proposed two types of arm movements (predictable and legible), and demonstrated that a legible arm movement, which is an overemphasised predictable motion of the human arm, can give more information about the action that the human or the robot is going to do. The experimental scenario involved two end-goals, close to each other. The participants were faster and more accurate to predict the end goal in the case of the overemphasised arm movement. However, there were only very few options in that scenario, and we argue that it would not generalize well if there were more end-goals (for example six as in our case).

We propose an alternative to embed action legibility with overemphasized arm motions, and extend the motion model to incorporate eye gaze information. Our approach improves legibility, by coordinating human-like eye-gaze behaviour with natural arm movements. The resulting robot's behaviour showed to be legible even for multiple sets of actions.

We validated these findings with a second human study, where subjects had to read/predict the intentions of a robot. In our experiments, it was much easier to read intentions of a robot than those of a human. We can explain this by looking at Fig. 6, that shows a side by side comparison of the action performed by the human and the robot. In the second pair of images, we see already a clear change in the eyes of the iCub, which is

not yet visible in the case of the human actor. This can be due to the high contrast between the white face and black eyes of the iCub. A different perspective on these results will be addressed in discussion of future work. A link for the video is provided here to illustrate the different steps taken in this work-https://youtu.be/HirRPgZGgFA.

The final conclusion taken from the second human study is the importance of the robot's gaze for the overall readability of the coordinated motion. Fig. 7 shows that just by looking at the arms without any gaze information the success rate drops below 85%. This also results in a slower prediction since the subjects have to wait for the arm of the robot to start moving which is slower than the movement of the eyes. Although 85% is a good result, it is only when we combine eyes and head movement that the results reach an almost perfect score. Our proposal combines the human gaze behaviour with the human arm movement to achieve legible behaviour to humans.

In the future we plan to improve our work in several ways, e.g. by expanding our dataset to more actors. We plan to revisit the modelling of the arm in order to better coordinate the overall eyes/head/arm speed. In our implementation, the robot arm controller is slower than the actual human arm motion. Moreover, while we carefully modelled the arm trajectories using GMMs, the gaze switching behaviour was not modelled with the same level of detail. While the robot gaze controller could qualitatively reproduce the human gaze-shift behaviours, "the human likeness" were not so close. We will thus investigate methodologies to model the gaze shift dynamics to a greater detail.

Our work stems the importance of non-verbal cues during a HHI, and the benefit of affording robots with the two-fold capacity: (i) interpreting those cues to read the action intentions of their human counterparts and (ii) to act in a way that is legible and predictable to humans.

ACKNOWLEDGMENT

We thank all colleagues that helped preparing and conducting the experiments, and all the people that participated in the human studies.

REFERENCES

[1] A. D. Dragan, K. C. T. Lee, and S. S. Srinivasa, "Legibility and predictability of robot motion," in *Proc. IEEE 8th ACM/IEEE Int. Conf. Human-Robot Interact.*, Mar. 2013, pp. 301–308.

[2] W. Erlhagen *et al.*, "Goal-directed imitation for robots: A bio-inspired approach to action understanding and skill learning," *Robot. Auton. Syst.*, vol. 54, no. 5, pp. 353–360, 2006.

[3] M. Huber, M. Rickert, A. Knoll, T. Brandt, and S. Glasauer, "Human-robot interaction in handing-over tasks," in *Proc. RO-MAN 17th IEEE Int. Symp. Robot Human Interact. Commun.*, Aug. 2008, pp. 107–112.

[4] L. Marin, J. Issartel, and T. Chaminade, "Interpersonal motor coordination: From human-human to human-robot interactions," *Interact. Stud.*, vol. 10, pp. 479–504, Dec. 2009.

[5] A. Sciutti, M. Mara, V. Tagliasco, and G. Sandini, "Humanizing human-robot interaction: On the importance of mutual understanding," *IEEE Technol. Soc. Mag.*, vol. 37, no. 1, pp. 22–29, Mar. 2018.

[6] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert, "Activity forecasting," in *Proc. Eur. Conf. Comput. Vis.*, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds., 2012, pp. 201–214.

[7] M. Pfeiffer, U. Schwesinger, H. Sommer, E. Galceran, and R. Siegwart, "Predicting actions to act predictably: Cooperative partial motion planning with maximum entropy models," in *Proc. IEEE /RSJ Int. Conf. Intell. Robots Syst.*, 2016, pp. 2096–2101.

[8] A. Sciutti, C. Ansuini, C. Becchio, and G. Sandini, "Investigating the ability to read others intentions using humanoid robots," *Frontiers Psychol.*, vol. 6, 2015, Art. no. 1362.

[9] J. Zhang, W. Li, P. O. Ogunbona, P. Wang, and C. Tang, "RGB-d-based action recognition datasets: A survey," *Pattern Recognit.*, vol. 60, pp. 86–105, 2016.

[10] H. S. Koppula and A. Saxena, "Anticipating human activities using object affordances for reactive robotic response," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 14–29, Jan. 2016.

[11] J. M. Gottwald, B. Elsner, and O. Pollatos, "Good is up-spatial metaphors in action observation," *Frontiers Psychol.*, vol. 6, 2015, Art. no. 1605.

[12] H. Admoni, A. Dragan, S. S. Srinivasa, and B. Scassellati, "Deliberate delays during robot-to-human handovers improve compliance with gaze communication," in *Proc. ACM/IEEE Int. Conf. Human-robot Interact.*, 2014, pp. 49–56.

[13] A. Fathi, X. Ren, and J. M. Rehg, "Learning to recognize objects in egocentric activities," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 3281–3288.

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

[15] M. A. Goodale, "Transforming vision into action," *Vis. Res.*, vol. 51, no. 13, pp. 1567–1587, 2011.

[16] T. Ohyama, W. L. Nores, M. Murphy, and M. D. Mauk, "What the cerebellum computes," *Trends Neurosci.*, vol. 26, no. 4, pp. 222–227, 2003.

[17] A. Shukla and A. Billard, "Coupled dynamical system based arm-hand grasping model for learning fast adaptation strategies," *Robot. Auton. Syst.*, vol. 60, no. 3, pp. 424–440, 2012.

[18] L. Lukic, "Visuomotor coordination in reach-to-grasp tasks: From humans to humanoids and vice versa," Ph.D. dissertation, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, Co-supervision with: Inst. Superior Técnico (IST) da Univ. de Lisboa, Lisbon, Portugal, p. 141, 2015. [Online]. Available: https://infoscience.epfl.ch/record/211059

[19] C. Elsner, M. Bakker, K. Rohlfing, and G. Gredebck, "Infants online perception of give-and-take interactions," *J. Exp. Child Psychol.*, vol. 126, pp. 280–294, 2014.

[20] M. Land, N. Mennie, and J. Rusted, "The roles of vision and eye movements in the control of activities of daily living," *Perception*, vol. 28, no. 11, pp. 1311–1328, 1999.

[21] M. Zheng, A. Moon, E. A. Croft, and M. Q.-H. Meng, "Impacts of robot head gaze on robot-to-human handovers," *Int. J. Social Robot.*, vol. 7, pp. 783–798, Nov. 2015.

[22] C. Pérez-D'Arpino and J. A. Shah, "Fast target prediction of human reaching motion for cooperative human-robot manipulation tasks using time series classification," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2015, pp. 6175–6182.

[23] S. Nikolaidis, R. Ramakrishnan, K. Gu, and J. Shah, "Efficient model learning from joint-action demonstrations for human-robot collaborative tasks," in *Proc. 10th Annu. ACM/IEEE Int. Conf. Human-Robot Interact.*, 2015, pp. 189–196.

[24] O. Palinko, F. Rea, G. Sandini, and A. Sciutti, "Robot reading human gaze: Why eye tracking is better than head tracking for human-robot collaboration," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2016, pp. 5048–5054.

[25] M. Kassner, W. Patera, and A. Bulling, "Pupil: An open source platform for pervasive eye tracking and mobile gaze-based interaction," in *Adjunct Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2014, pp. 1151–1160.

[26] C. Kothe, "Lab streaming layer (LSL)," [Online]. Available: https://github.com/sccn/labstreaminglayer, Accessed on: Feb. 2015.

[27] D. C. Montgomery and G. C. Runger, *Applied Statistics and Probability for Engineers*. New York, NY, USA: Wiley, 2010.

[28] S. Calinon, F. Guenter, and A. Billard, "On learning, representing and generalizing a task in a humanoid robot," *IEEE Trans. Syst. Man Cybern. B*, vol. 37, no. 2, pp. 286–298, Apr. 2007.

[29] A. Moon *et al.*, "Meet me where i'm gazing: How shared attention gaze affects human-robot handover timing," in *Proc. ACM/IEEE Int. Conf. Human-robot Interact.*, 2014, pp. 334–341.

[30] A. Roncone, U. Pattacini, G. Metta, and L. Natale, "A cartesian 6-DOF gaze controller for humanoid robots," in *Proc. Robot.: Sci. Syst.*, 2016.