Action Anticipation: Reading the Intentions of Humans and Robots

Nuno Duarte[1], Jovica Tasevski[2], Moreno Coco[3], Mirko Raković[1,2] and José Santos-Victor[1]

***Abstract*— Humans have the fascinating capacity to understand, and anticipate the actions of other humans in the same space, without requiring verbal communication. This "intention reading" capacity is underpinned by a common motor-repertoire that is shared by all humans, and afforded by a subtle coordination of eye-head-arm movements that encodes the cues and signals that will ultimately be deciphered by another human.**

In this paper we study the action anticipation capacity of humans and robots with a focus on three steps: (i) conducting human interaction studies to record a set of relevant motion signals and cues, (ii) use these data to build computational motor control models, and (iii) incorporate this model in a robot controller.

In the human studies we ask participants to guess what action the actor was performing: *giving* or *placing* an object on a table. Our results reveal that *giving* actions present a more complex behavior of the human gaze compared to the *placing* actions. These findings are integrated in our motor control together with the arm movement modeled from human behavior. Gaussian Mixture models the human arm movement, and then Gaussian Mixture Regression generates the controller. The readability of the controller is tested on a human-robot scenario validating the results acquired from the human experiment.

Our work is a step forward to building robotic systems that are not only capable of reading and anticipating the actions of human collaborators but also, at the same time, to act in a way that is legible to their human counterparts.

I. Introduction

Humans often have to share a workspace in different manufacturing and domestic environments and have to infer each other's action so to optimize their coordinated interaction. In order for us to have a successful human-robot interaction (HRI) in such environments, we must understand the intention of the robot so as to correctly anticipate its action. Our goal in this paper is to use human bio-functionally inspired action representation as the robot's motor control to anticipate the robot's action.

In recent years, research as focused on the human part of the equation ([1], [2], [3]) in order to develop better cognitive systems that are able to cooperate in dyadic actions such as handover of objects (Fig. 1). The improvement of humanoid robots, which have similar motor system, allows for human-like interaction which in turn facilitates the cooperation in joint actions. This idea is corroborated with the discovery of mirror neurons [4], in monkeys and later in humans. This neurons activate in the premotor cortex when executing an action and when we recognize the action that is executed. Neuroscientists believe that the mirror neuron system is a way of understanding the actions and intentions of other humans This supports the assumption that humans will understand much more easily the actions of a humanoid robot. Moreover, in [5] we see that there is a preferential bias towards familiar verbal and non-verbal cues when communicating with robots.

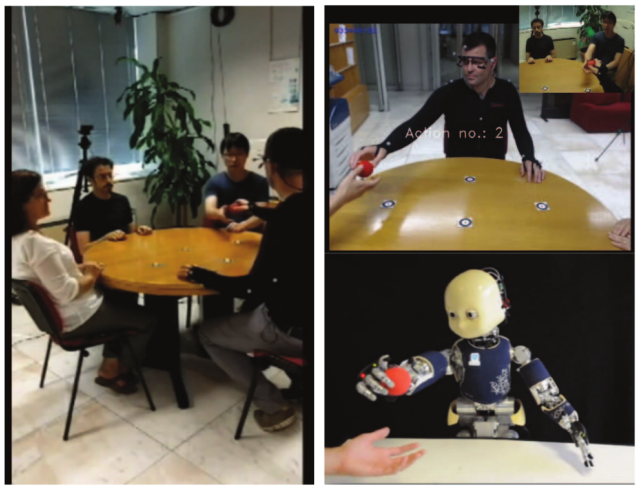


Fig. 1: Human-Human Interaction: an experiment involving one actor (top-right) *giving* and *placing* objects and three subjects reading the intentions of the actor (left); Human-Robot Interaction: a robot performing the human-like action and subjects try to anticipate the robots intention (bottom-right).

For the human study, we design an experiment a human-human interaction (HHI) where one actor is interacting with three other people in a dyadic action. During the interaction we are analyzing the actor's body and eye movements in a *giving* and *placing* actions. This analyses is achieved with the use of a motion capture system (OptiTrack) for the body motion, and an eye tracking system (Pupil Labs) for tracking the gaze fixation. With the recordings we create video fractions for each action in order to ask participants the intention of the actor at different moments in time. Since the focus is on non-verbal communication, we fraction the videos in 4 different triggers: gaze shift, head shift, arm acceleration, and arm deceleration. We see that with just the gaze information 85% of participants guess correctly that the human was executing a *placing*. For the *giving* action we get

*This work was supported by EU H2020 project 752611 - ACTICIPATE, FCT project UID/EEA/50009/2013 and RBCog-Lab research infrastructure. We would like to thank all of our colleagues that supported us in preparing and conducting the experiments, and everyone that participated in them.

[1]N. Duarte, M. Raković and J. Santos-Victor are with the Vislab, Institute for Systems and Robots, Instituto Superior Técnico, University Lisbon, Portugal, {nferreiraduarte, rakovicm, jasv}@isr.tecnico.ulisboa.pt

[2]J. Tasevski is with the Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia, tasevski@uns.ac.rs

[3]Moreno Coco is with the Department of Psychology (Centre for Cognitive Ageing and Cognitive Epidemiology), University of Edinburgh, Scotland, moreno.coco@ed.ac.uk

unexpected results that are analyzed in detail in section IV-C. We show that we do not get a significant improvement of the prediction at every different fraction of the videos. Only the first three triggers of information give relevant information to decide the intention of the actor.

The second part of the paper is on using the dataset generated to build a model of the human legible motion and develop a controller of that model for the humanoid robot iCub. We show that human gaze behavior in a robot improves the prediction of intention from 80% to 90%. The participants are able to predict quicker and more correctly the action of the robot when the gaze is visible. Our experiments also show that the arm trajectory of the robot is legible to humans although not as important as the gaze information. The paper concludes with final discussion in Section VII.

II. State of the Art and Contribution

For developing the action anticipation algorithms, researchers are relying on different datasets. Zhang et al. [6] presents a survey on RGB-D based action recognition datasets. In [7], authors are using CAD 120 dataset to develop jointly identifying human sub-actions to obtain a clear labeling of the activities being performed. For human study research [8] and HRI where gaze tracking is significant [9], researchers are relying on gathering their own or using some of the existing datasets, such as GTEA dataset [10]. However, there is a lack of synchronized and labeled video+gaze body motion datasets. In this paper it will be presented a procedure for collecting one such dataset[1].

The goal for preparing the dataset was to support the development of bio-inspired coordinated gaze and body motion controller. Research in neurobiology provides several well-known models concerning the gaze, hand and gaze-arm coupling. One group of researchers are focused on investigating cortical structures such as the posterior parietal cortex, the premotor and the motor cortices [11], [12]. Another stream of research interest has been focused on modeling the role of the cerebellum in the motor loop, movement generation and synchronization of sensory-motor system [13], [14]. These findings are used by [15], [16] to develop coupled dynamical systems framework for arm-hand and eye-arm-hand coordinated motion control for robots based on the measures in the human study. The framework provides fast and synchronous control of the eyes, the arm and the hand, mimicking similar control system found in humans. The framework is focused on intra-personal motor control coupling, whereas in this research interest is more in the coordination of sensory-motor system for interpersonal coupling.

Authors in [17] presented the study with infants to investigate the perception during object handover. The eye movements of 12-month-olds were recorded as they observed a give-and-take interaction. Their findings provide evidence that properties of social action goals influence infants' online gaze during action observation. On the other hand, an individual action such as making tea [18], has a specific task order and so it requires multiple gaze fixations to complete the action. However, this research just focuses on the fixation of either a particular object or hand, not at both during the same action. The work by Meng et al [19] addresses this peculiarity of the human gaze in a dyadic action. They build an experiment where different types of gaze trajectories are tested in a human-robot scenario. It shows that humans prefer when the robot fixates the person's face and then switches to the handover position than just looking either the face or the handover position exclusively. Although this paper proves there is a contextual behavior of the gaze, the human-robot study performed has some limitations. It is stated that the robot switching behavior of the gaze is actually head orientation. This is due to the robot's nonhuman like head not having degrees of motion of the eyes. In our work we will use an eye tracking device to observe the actual gaze fixation points during the interaction as this provides better accuracy than just the head orientation [20]. Also, in the experiment all the subjects were already aware of the type of action the robot was going to perform. There is already a bias towards handover even before the action is executed.

Dragan et al. [21] discusses the terminology of predictable and legible trajectories of motion. It proves that robots that have legible motions their intention can be guessed sooner in the movement of the arm. In our work we want to follow this concept and extend it to the gaze behavior of the robot. Our proposal is that legibility improves the action intention of the robot and so a legible gaze behavior will improve the overall understanding of the robot's action. We believe that to achieve a legible gaze behavior we would have to study the humans' eyes during interpersonal interactions. Moreover, for unbiased behavior, this experiment needs to have distinct actions of different end-goals such as intrapersonal actions (*placing*) and dyadic actions (*giving*).

Thus, the summarized contribution of this paper is:

(i) a publicly available dataset with synchronized videos, gaze and body motion,
(ii) robots with human gaze and arm motion behavior are more legible to humans than with just one or the other.
(iii) discovering the contextual relevance of gaze fixations cues during a *giving* action

III. Human Study

In this section we perform a human-human study to analise human behavior in interpersonal interactions involving *placing* and *giving* objects. During this study we collect the body and eye's movement of the actor who is executing the aforementioned actions.

A. Experiment Description

The experiment setup is illustrated in Fig. 2. For each experiment trial there is one actor executing the actions *placing* and *giving* an object, and three subjects that the actor interacts by either *placing* in front of one of the subjects or

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

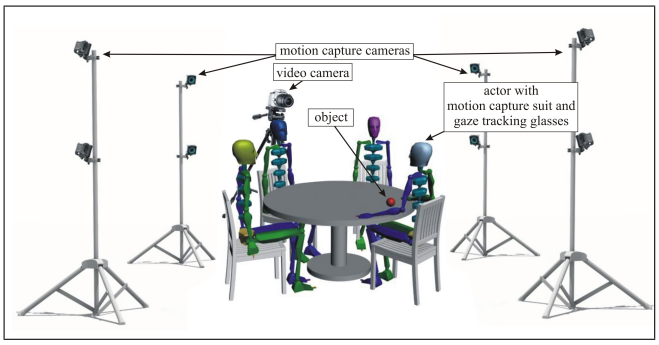


Fig. 2: Experimental setup to record synchronized video, body motion and eye gaze during the actions of *placing* the object (red ball) on the table and *giving* the object to the humans across the table.

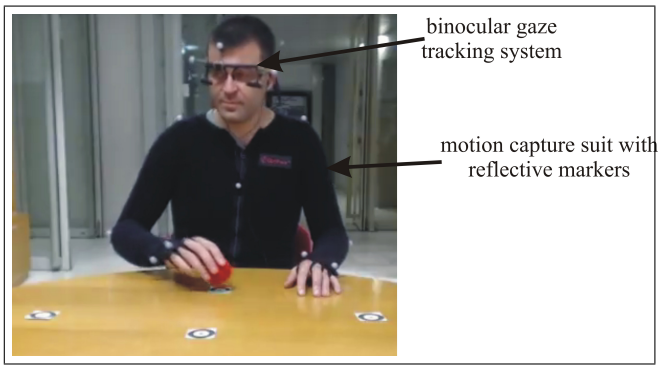


Fig. 3: Snapshot from a second video camera that is recording an actor. There are 4 markers visible in the image. Representing the three *placing* positions of each subject, and the initial position where the object starts from to initiate a new action.

giving the object to one of the three. The actor would grab the object from the initial position (3) and execute one of the preselected actions:

- ***placing*** on the table to the actor's **left**: p(L)
- ***giving*** the ball to the person on actor's **left**: g(L)
- ***placing*** on the **middle** of table: p(M)
- ***giving*** to the person in the **middle**: g(M)
- ***placing*** on the table to the actor's **right**: p(R)
- ***giving*** to the person on actor's **right**: g(R)

The actions 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 are not predetermined but rather randomly selected to prevent the actor from adapting its posture prior to initiation. Every action begins and ends with the actor placing the ball back to the initial position on the table.

B. Hardware and Software Setup

OptiTrack motion capture system records the movements of the actor, consisting of 12 motion capture (MoCap) cameras all around the environment and a motion capture suit with 25 markers on the upper torso, arms, and head worn by the actor. The data provided by the MoCap is 3d Cartesian coordinates of every body part (head, torso, right-arm, left-arm). For the gaze recording, we used the binocular Pupil Labs eye tracker [22]. It is a mobile eye tracking headset for tracking the actor's fixation point. The head markers were placed on the Pupil Lab tracker to establish relationship between the gaze and body motion reference coordinate frames. To record the scene, three video cameras are placed at very specific points to provide different angles that will complement during the evaluation phase. The first camera is the worldview perspective of the actor from the Pupil Labs eye tracking headset (Fig. 1 (top-right) 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. 3). It's from this perspective that we take the recordings to build an experiment to test the action understanding in a HHI. One of the problems that we encountered from this view is that the eye tracking system sometimes would partially occlude the eyes, which makes it harder for humans to see the gaze trajectory when trying to understand the intention of the actor. The third camera was located further from the environment, looking inwards, giving a proper reading of the subject's actions and a different angle of the subject's arm movements (Fig. 1 (left)). After the hardware was all in place, the next step was to run the needed software and start the recordings.

To collect all the sensory information, OptiTrack's Motive and Pupil Lab's Pupil Capture software were used. The pupil detection algorithm uses edge detection in order to find the dark pupil in the infrared (IR) illuminated eye camera image [23]. Before the beginning of the recordings, both sensors were calibrated.

The video cameras do the recording of the experiment from different perspectives: view of the actor, the top view of the table, and view of the whole experiment. The view of the actor is recorded when the Pupil Capture system from Pupil Labs is recording. The top view and the view of the whole experiment are two distinct video cameras that need to be started individually.

C. Synchronized Body Motion and Gaze Dataset

The total of 120 trials are performed with actions p(L), g(L), p(M), g(L), p(R), g(R), performed 20, 17, 23, 19, 17 and 24 times respectively. The Pupil Lab binocular gaze tracking system recorded world camera video and gazed data in 3D Cartesian coordinates at 60Hz, the motion capture system recorded the movements of the body at 120Hz, and video camera facing the actor, recorded video at 30Hz. The data from all sensing systems were gathered at one place keeping the information of the timestamps of each of the sensing system and extending the data with the internal clock information that is used as a reference to synchronize the sensory information.

The collected data in this research are used for two purposes. The recordings of the camera facing the actor are used to prepare a questionnaire for understanding how different cues influence human's decision on the action to be performed. The body motion and eye gaze are used for further data processing, motion modeling and implementa-

tion of the bio-inspired motion controller in order to achieve human-like behavior.

IV. A Study of Human Action Anticipation

To better understand what cues the humans use to anticipate other human's action, and how the cues are related to the spatial distribution action we prepared a questionnaire related to the actions performed by an actor.

A. Participants

The experiment was conducted with fifty five participants (40 male, and 15 female) age 31.9±13 (mean±SD). There were thirteen teenagers and six people over fifty years of age. Three of the subjects were left-handed, approximately 62% were students, 27% were professors, 7% were researchers, and 4% were staff members. All subjects were naive to the purpose of the experiment.

B. Task

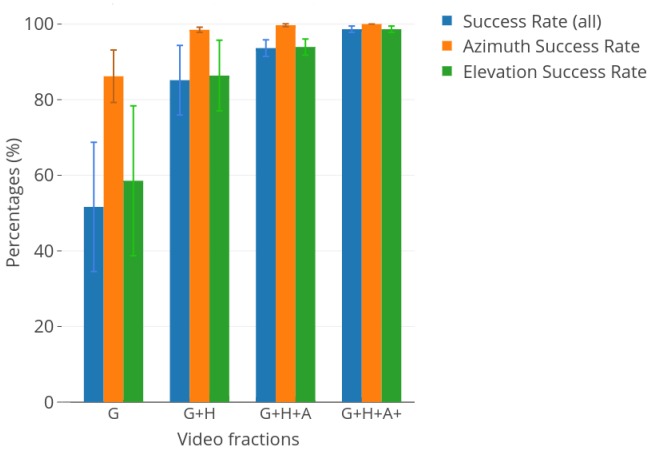
The task of the experiment consist of a questionnaire with 24 questions[2]. Before the question is show to the subject they have to watch a short video the actor performing one of the six possible actions. Based on the video shown, the participant had to identify what is the intended action of the actor. The videos were fractioned into 4 sizes according to the actor's behavior. This video fractions can be comprehended as short videos beginning when the actor grabs the object and ending until a certain moment in time where new information is provided. The ending of the short videos can be in 4 different time instances:

- G - when there is a saccadic movement of the eyes towards the intended goal
- G+H - 'G' happens plus the head orientates towards the same goal
- G+H+A - 'G' plus 'H' happens plus the arm starts moving towards the goal
- G+H+A+ - 'G' plus 'H' plus 'A' plus the arm finishes the trajectory towards the goal

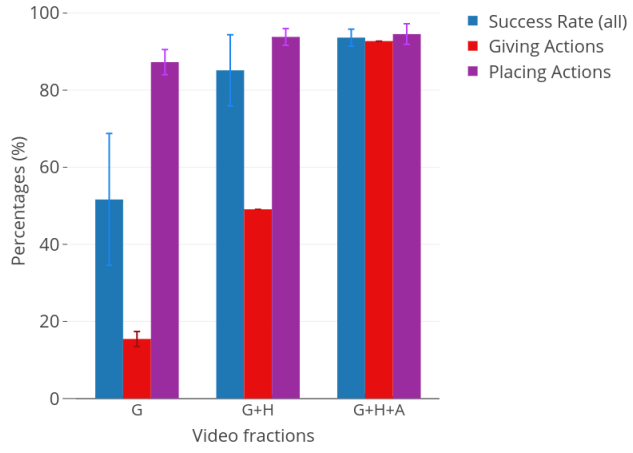
The last group of videos (i.e., G+H+A+) was used as a golden standard to remove outliers. Out of 24 videos, first three are used to familiarize participants with the questionnaire and are neglected for further analysis. Out of remaining 21 questions, five questions are from the G difficulty level, six 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

Fig. 4a shows the overall success rate of the participants (blue) in giving the correct answer in one out of six actions to be performed by an actor. It is clear that the average success rate for each group of video fractions is increasing as more information is provided to the subject, whereas standard deviation decreases, stating that participants are more and more similar in the correctness rate. The analysis is further refined by looking at the success rate across the spatial directions, azimuth and elevation Fig. 4a, orange and green lines, respectively. We notice that for azimuth direction the more information provided does not improve the legibility of the action. With just gaze information we get an average of 85% of success guessing and with head orientation it reaches almost 100% of correct guesses. This means that for legible motions with human gaze behavior the action is anticipated without the need of arm movement. For the elevation direction, corresponding to *placing* and *giving* action, we get different results. To get a better analyses of the results we separate into two independent columns, red and purple lines, respectively (Fig. 4b). Here, we notice a clear trend in the values. For *placing* actions the conclusions are identical to what was previously discussed. As for the *giving* action a more in depth analyses is required to understand the values obtained.



(a)



(b)

Fig. 4: 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.

A two-way ANOVA [24] was used to analyze the answers

[2]The questionnaire can be seen at the following web address: https://sites.google.com/site/acticipatequestionnaireno01/

of the *giving* actions. Binary variables such as, the amount of information given to the subjects, and the type of actions are the factors studied. The main effect of the amount of information given to the subject, so the type of video fraction, was also significant on the subject's success, F(2,5560)=1396.76, p<0.0001. Also, these effects shows significant interaction between the type of the action and the amount of information given to the subject, F(2,5560)=537.70, p<0.0001. This validates that success of guessing the correct answer depends on the type of action (*giving* and *placing*) and also on the amount of information ('G', 'G+H', 'G+H+A', 'G+H+A+'). For the *giving* actions we notice that subjects have success rate lower than if they picked randomly. Meaning that there is a bias towards answering *placing* when it is a *giving* action for the video fractions of 'G' and 'G+H'. The main effect of the type of action on subject success was significant, F(1,5560)=2306.78, p<0.0001. This confirms it is easier to recognize the action *placing* compared to the action *giving*.

With this experiments we get a notion about the importance of gaze in a dyadic action. In a HHI gaze information provided by the eye movements gives the necessary information to predict the intention of the other subject. For the *giving* actions this is not the case but we believe that it is the experiment geometry setup that gives a unintentional bias towards the action that requires the least energy, *placing* the object on the table. Without this constraint we can conclude that human gaze provides information to correctly guess the action and so extending the legibility of the robot to include a human-like gaze will improve the interaction in human-robot scenario.

V. Modeling Human Motor Control

To model the action of actor we use a Gaussian Mixture Model (GMM) [25] to encode 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 of arm trajectories for all actions, and ξ_j is the Cartesian coordinates of the hand for a *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.

$$p(k) = \pi_k \tag{1}$$

$$p(\xi_j|k) = \mathcal{N}(\xi_j; \mu_k, \Sigma_k) \tag{2}$$

$$= \frac{1}{\sqrt{(2\pi)^D|\Sigma_k|}} e^{-\frac{1}{2}((\xi_j-\mu_j)^T\Sigma_k^{-1}(\xi_j-\mu_j))} \tag{3}$$

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. 5 shows an example of the recorded trajectories of the actor's hand during execution of the p(R) action. In the middle column the recorded trajectories are encoded in GMM, whose covariances matrices is represented by ellipses. It is used 4 Gaussian distributions to model the behavior of the arm trajectory for each Cartesian coordinate. This option was chosen taking 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 follows in [25]. Note that this is equivalent to averaging over time the mean value and covariance as it would produce unnatural motions and it would increase dramatically the dimension of parameters. The right column represents the GMR output of the signals in bold, and the covariance information as the envelope around the bold line.

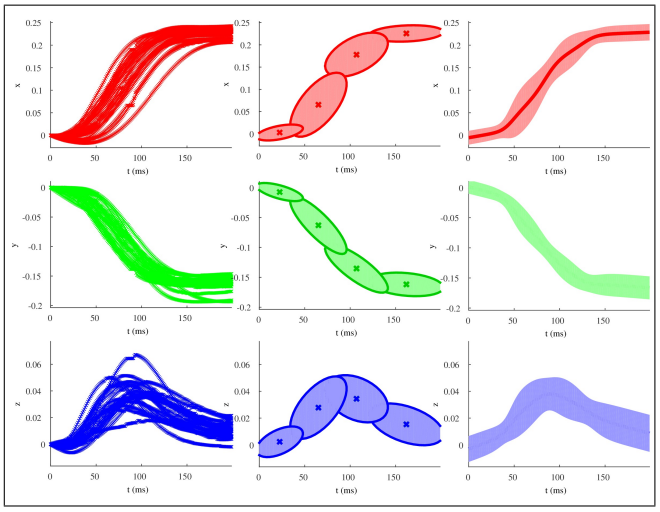


Fig. 5: Recorded coordinates of human hand performing p(R) action, representation of corresponding covariance matrices and output from GMR with covariance information.

The same modeling is done for all six actions. Fig. 6 shows spatial distribution of the recorded data for all six actions represented by six different colors. In Fig. 6 bellow human motion recording, a sequence of temporal values is used as query points to retrieve expected spatial distribution through GMR.

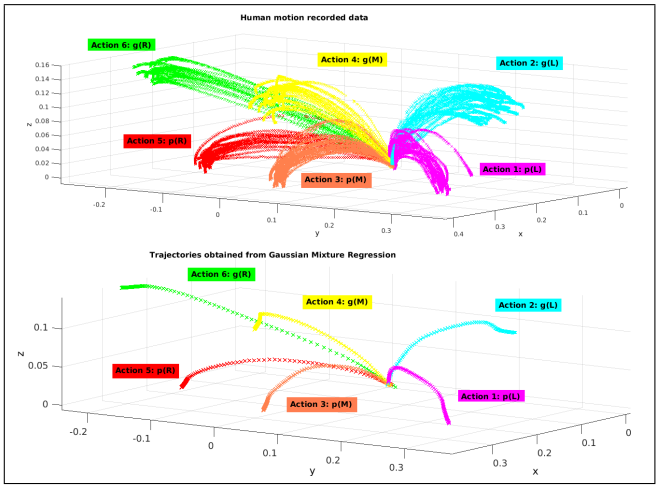


Fig. 6: Spatial distribution of hand motion for all six actions (top) and corresponding output from GMR (bottom)

Our initial thought was to use the same approach for estimating a time-invariant spatial distribution of fixation

(a) (b) (c) (d) (e)

Fig. 7: The 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.

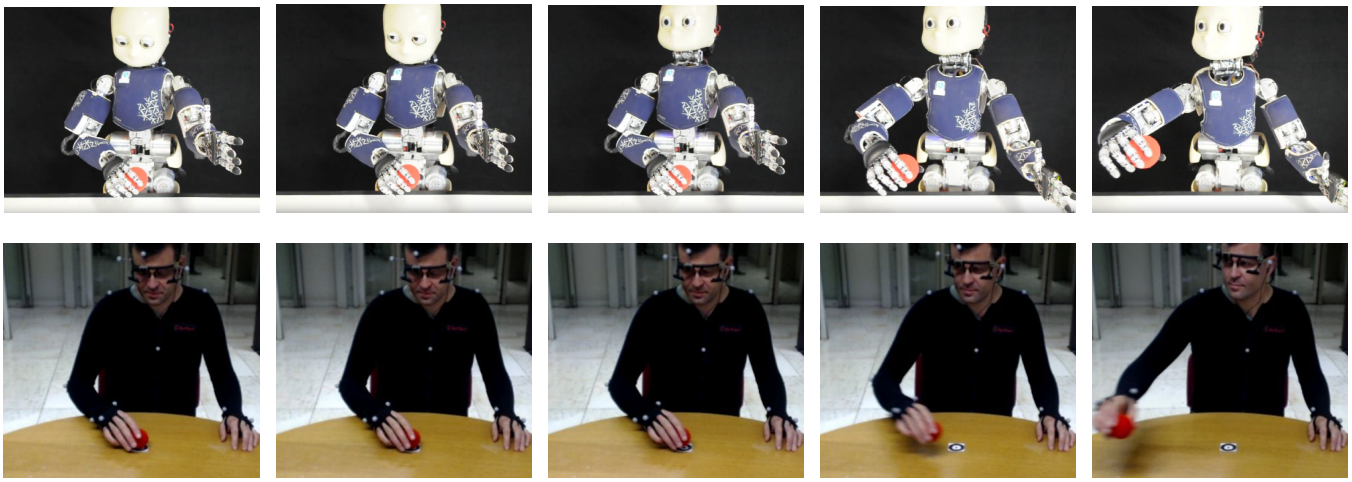


Fig. 8: The sequence of images of a robot (top) and an actor (bottom) performing the g(R) action. The first sequence is the initial point for both the actor and the robot. The second stage corresponds to when the short video stops at the video fraction 'G'. The third is at video fraction 'G+H'. Forth and fifth sequences are for the final two video fractions, corresponding to the arm motion.

point for driving eye-head coupled controller on iCub robot [26]. Though, from the authors in [27] it is known that in human-human interactions, during *giving* actions, humans vary their gaze behavior. The HRI handover speed improves when the robot looks at the handover position (which in the paper is called "projected handover"). When the robot performs the switching behavior of looking at the human's face and then to the handover position it does not improve the speed of the human reaching time but it does improve the perception of the interaction [19]. However, in that human study the experiment involved two humans *giving* a bottle to each other. So, throughout the experiment they were always expecting a *giving* action. In this situations it would not be an unbiased human gaze behavior as if the human were unaware of the intention of the other. In our experiment, the actor was instructed the next action (randomly selected out of six possible) right after the previous one had finished. Since there were three *giving* actions, and three *placing* actions, we could expect a natural gaze behavior of the actor in every case. They concluded that the most dominant gaze behavior was the 'Shared Attention Gaze'. Since the experiment is not entirely unbiased, their conclusions need to be validated. Moreover, they characterize 'Shared Attention Gaze' as "consists of the giver gazing at a projected handover location as s/he reaches out to execute the handover", and mention that "there is no eye contact between the giver and the receiver throughout this handover gaze pattern". This has nothing to do with sharing gaze attention as they never gaze each other in the process.

Another limitation found in Moon, et al [27] is the fact that the analyses of the different gaze behaviors were done empirically (manually looking at the videos). From our dataset were are able to detect the fixation points, the actual points of interest in a handover task, and the duration of gaze between each switching behavior. This information will provides in the future to build a biologically inspired human eye controller.

In this paper, our contribution to the study of human gaze behavior is an in depth analyses of the human gaze with recourse to an eye tracking system (Pupil Labs). Fig. 7 shows five different cases of the spatiotemporal distribution of the fixation point marked with a green circle. In Fig. 7a shows the spatiotemporal distribution of fixation points for the p(M) *placing* action in which the green circle is always concentrated around the goal position of the red ball. Fig. 7b-7e are showing the spatiotemporal distributions of 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. Having this brighten up, we designed coordinated spatiotemporal change of the fixation point.

In the case of the *placing* actions, represented in Fig. 7a,

the fixation point is at the initial position of the ball. When the movement starts, the fixation point changes to the end-goal ("projected handover") position of the ball and remains there until the end of the movement. For the *giving* actions, in the beginning, the fixation point is also set to be equal to the initial position of the ball. Once the action starts, the fixation point can be set either to a lower position, representing the hand, or higher position, representing the face of the human it will interact. The fixation point is fed to the coupled eye-head controller that executes saccadic eyes movement followed by further coordinated motion of the eye and neck joints. The sequence of images showing the execution of g(R) action with the iCub robot and corresponding images of the actor are given in Fig. 8 for the case when fixation point is first set to lower position and then switched to higher position.

During the experiments for the same actor and same action all 4 possible gaze behaviors were occurring. No instructions to the actor were given regarding his gaze behavior. We agree with the authors in [19] that the gaze behavior that involves a switching of the gaze attention between the face of the human and the handover location or hand might be the best option for HRI handover. Nevertheless, further research is required to prove this. As future work we are developing experiments with humans and robots that test the perception time of the human to a robot handover with gaze behavior as the variable. For now, to validate the legibility of the human motion controller, we implement on the robot the model of the human arm motion with the eye-head human gaze behavior using the Cartesian 6-DOF gaze controller in [26].

VI. Closing the Loop: Anticipating the Robot's Intention

To validate the legibility of robot's intention we prepared a second questionnaire[3] with the same set of six actions, but now performed by a robot. Initially, the same set of questions with four different video fractions was being prepared. When reviewing the questionnaire it was evident that the iCub's eye movements are unoccluded by eyelids or even the eye tracker. Even for the video fractions 'G' it was quite simple to guess. For that reason, we considered two additional conditions. Blurring the video occlude the eyes or the entire head. This is to reinforce the importance of gaze information in understanding the intention of the robot. Our experiment showed that difficulty increases with the size of the blurred area.

The human study involved 20 participants answering 36 question: 18 without any blurring, 12 with eye blurring, and 6 with the whole head blurred. The format of the case study was the same as the previous one, but now with the robot performing the actions. Fig. 9a shows the overall success rate of the participants in identifying the action performed by a robot as well as if it was a *giving* or *placing* action.

As in the first case study, the average success rate increases as more information is provided to the participants. Fig. 9

[3]The second questionnaire can be seen at the following web address: https://sites.google.com/site/acticipatequestionnaireno2/

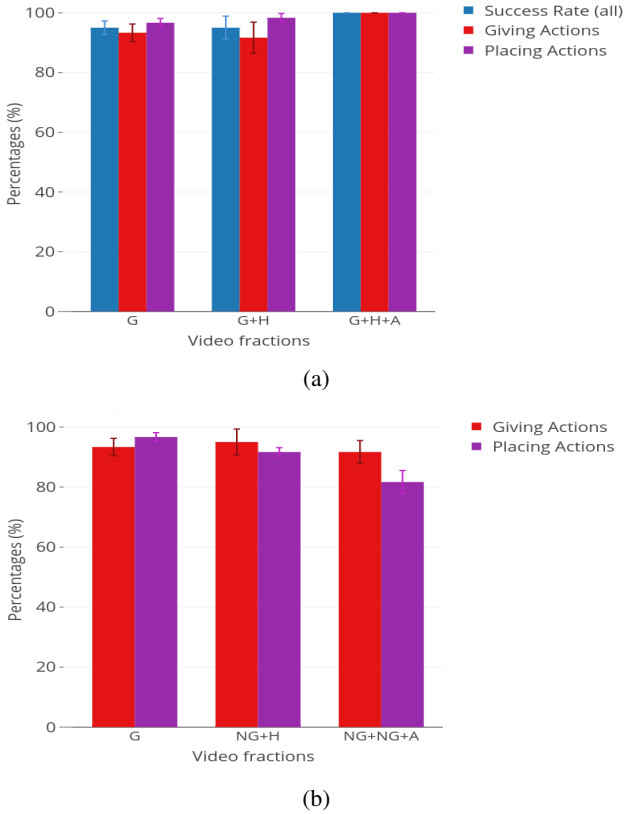


Fig. 9: The success of the participants identifying the correct action a) overall success rate; b) success rate in identifying the *giving* and *placing* actions.

shows the success rate across different blurring levels for *placing* and *giving* action. We can see that when adding blurring to the eyes with just the head information ('NG+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 guess the intention for the robot. When the blurring involves the whole head, the only information provided is from the legible motion of the arm. In [21] paper we know that this is a predictable arm motion and will not give the most information to the user.

The conclusions after this HRI experiment is that the action of the robot becomes much more legible with the integration of gaze to the controller. Although, the arm motion is not legible in the terminology developed by Anca Dragan et al, when added human gaze behavior we can have much better results and in end more biologically inspired, e.g. natural behavior.

VII. Discussions and Conclusion

In this research, we collected and made it publicly available the dataset of synchronized videos, gaze and body motion data. The data recordings are utilized in two ways. First, the video material is used to prepare the questionnaire for studying anticipation of other humans' actions. Second, the motion data is used to construct a bio-inspired motion

controller. For modeling, we used GMM and utilized GMR to generate trajectories to drive iCub's arm.

The readability of robot's intention showed to be at the level comparable to human's. There is a certain difference in results of two studies when we are looking at the plots for the success rate in identifying the *placing* and *giving* success rate, as well as in identification of the direction of the action. This can be partially explained by different kinematic and dynamic properties between human's and the robot's, but the overall results show a clear and similar increase in the success rate across different difficulty levels.

There are several limitations in this work. For starters, the experiments were performed with just one actor. It would be better if this experiments were done with more actors in order to generalize the analyses from this studies.

REFERENCES

[1] W. Erlhagen, A. Mukovskiy, E. Bicho, G. Panin, C. Kiss, A. Knoll, H. Van Schie, and H. Bekkering, "Goal-directed imitation for robots: A bio-inspired approach to action understanding and skill learning," *Robotics and autonomous systems*, vol. 54, no. 5, pp. 353–360, 2006.

[2] M. Huber, M. Rickert, A. Knoll, T. Brandt, and S. Glasauer, "Human-robot interaction in handing-over tasks," in *Robot and Human Interactive Communication, 2008. RO-MAN 2008. The 17th IEEE International Symposium on*, pp. 107–112, IEEE, 2008.

[3] L. Marin, J. Issartel, and T. Chaminade, "Interpersonal motor coordination: From human–human to human–robot interactions," *Interaction Studies*, vol. 10, no. 3, pp. 479–504, 2009.

[4] G. Rizzolatti, L. Fadiga, V. Gallese, and L. Fogassi, "Premotor cortex and the recognition of motor actions," *Cognitive brain research*, vol. 3, no. 2, pp. 131–141, 1996.

[5] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics and autonomous systems*, vol. 42, no. 3, pp. 143–166, 2003.

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

[7] H. S. Koppula and A. Saxena, "Anticipating human activities using object affordances for reactive robotic response," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 14–29, 2016.

[8] J. M. Gottwald, B. Elsner, and O. Pollatos, "Good is upspatial metaphors in action observation," *Frontiers in psychology*, vol. 6, 2015.

[9] H. Admoni, A. Dragan, S. S. Srinivasa, and B. Scassellati, "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, ACM, 2014.

[10] A. Fathi, X. Ren, and J. M. Rehg, "Learning to recognize objects in egocentric activities," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference On*, pp. 3281–3288, IEEE, 2011.

[11] S. M. Beurze, I. Toni, L. Pisella, and W. P. Medendorp, "Reference frames for reach planning in human parietofrontal cortex," *Journal of neurophysiology*, vol. 104, no. 3, pp. 1736–1745, 2010.

[12] M. A. Goodale, "Transforming vision into action," *Vision research*, vol. 51, no. 13, pp. 1567–1587, 2011.

[13] R. Miall, G. Reckess, and H. Imamizu, "The cerebellum coordinates eye and hand tracking movements," *Nature neuroscience*, vol. 4, no. 6, p. 638, 2001.

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

[15] A. Shukla and A. Billard, "Coupled dynamical system based arm–hand grasping model for learning fast adaptation strategies," *Robotics and Autonomous Systems*, vol. 60, no. 3, pp. 424–440, 2012.

[16] L. Lukic, "Visuomotor coordination in reach-to-grasp tasks: From humans to humanoids and vice versa," PhD Thesis, 2015.

[17] C. Elsner, M. Bakker, K. Rohlfing, and G. Gredebäck, "Infants' online perception of give-and-take interactions," *Journal of experimental child psychology*, vol. 126, pp. 280–294, 2014.

[18] 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.

[19] M. Zheng, A. Moon, E. A. Croft, and M. Q.-H. Meng, "Impacts of robot head gaze on robot-to-human handovers," *International Journal of Social Robotics*, vol. 7, no. 5, pp. 783–798, 2015.

[20] 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 *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*, pp. 5048–5054, IEEE, 2016.

[21] A. D. Dragan, K. C. Lee, and S. S. Srinivasa, "Legibility and predictability of robot motion," in *Human-Robot Interaction (HRI), 2013 8th ACM/IEEE International Conference on*, pp. 301–308, IEEE, 2013.

[22] M. Kassner, W. Patera, and A. Bulling, "Pupil: An open source platform for pervasive eye tracking and mobile gaze-based interaction," in *Adjunct Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '14 Adjunct, (New York, NY, USA), pp. 1151–1160, ACM, 2014.

[23] L. Świrski, A. Bulling, and N. Dodgson, "Robust real-time pupil tracking in highly off-axis images," in *Proceedings of the Symposium on Eye Tracking Research and Applications*, pp. 173–176, ACM, 2012.

[24] D. C. Montgomery and G. C. Runger, *Applied statistics and probability for engineers*. John Wiley & Sons, 2010.

[25] S. Calinon, F. Guenter, and A. Billard, "On learning, representing and generalizing a task in a humanoid robot," *IEEE Transactions on Systems, Man and Cybernetics, Part B*, vol. 37, no. 2, pp. 286–298, 2007.

[26] A. Roncone, U. Pattacini, G. Metta, and L. Natale, "A cartesian 6-dof gaze controller for humanoid robots.," in *Robotics: Science and Systems*, 2016.

[27] A. Moon, D. M. Troniak, B. Gleeson, M. K. Pan, M. Zheng, B. A. Blumer, K. MacLean, and E. A. Croft, "Meet me where i'm gazing: how shared attention gaze affects human-robot handover timing," in *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pp. 334–341, ACM, 2014.