Gaze-based intention recognition for pick-and-place tasks in shared autonomy

Stefan Fuchs[1] and Anna Belardinelli[1]

***Abstract*— Shared autonomy aims at combining robotic and human control in the execution of remote, teleoperated tasks. This cooperative interaction cannot be brought about without the robot first recognizing the current human intention in a fast and reliable way, so that a suitable assisting plan can be quickly instantiated and executed. Eye movements have long been known to be highly predictive of the cognitive agenda unfolding during manual tasks and constitute, hence, the earliest and most reliable behavioral cues for intention estimation.**

In this study, we present an experiment aimed at analyzing human behavior in simple teleoperated pick-and-place tasks in a simulated scenario and at devising a suitable model for early estimation of the current proximal intention, that is either the reaching target or the place-down location. We show that scan paths are, as expected, heavily shaped by the current intention and that a Gaussian Hidden Markov Model achieves a good prediction performance, while also generalizing to a new object configuration and new users. We finally discuss how behavioral and model results suggest that eye movements reflect to some extent the invariance and generality of higher level planning across object configurations.

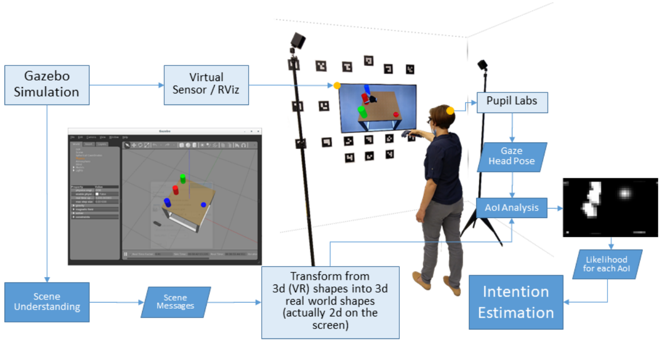


Fig. 1: The experiment was carried out in a simulated scenario. The Gazebo simulator features physics and a virtual sensor, with its measurements shown on the screen. The user controls the Pick-and-Place-Task with a HTC Vive controller, tracked by the Vive Lighthouse system. At the same time, the Vive controller and a Pupil Labs eye-tracking system are used to estimate the human state.

I. INTRODUCTION

Shared autonomy has recently emerged as an ideal trade off between full autonomy and complete teleoperation in the execution of remote tasks. The benefits of this approach rely on assigning to each party the aspects of the task for which they are better suited. The lower kinematic aspects of action execution are usually left to the robot while higher level cognitive skills, like task planning and handling unexpected events, are typically concurrently exercised by the human, in a blend that can entail different degrees of autonomy for the robotic part ([1], [2], [3]). Considering the often large asymmetry in terms of degrees of freedom or kinematic capabilities between the user input controller (e.g. joysticks) and the robotic effector, shared autonomy eases the operator cognitive load and speeds up execution. Since the user is setting the goals and the ways to achieve them, this collaborative effort relies on the robotic partner to first recognize the current human intention (*intent recognition*) and only afterwards to decide how much to assist with the execution (*arbitration*). Intention recognition should thus happen as early and as naturally as possible in order for the user to be relieved of explicitly directing the robot and for the robot to timely initiate the assisting action. To this end, although a number of approaches have been proposed that rely on intent recognition from the user control input driving the robotic movement ([4], [5], [6], [7], [8], [9]), the most natural and timely way to predict intention both in assistive technologies and remote manipulation is certainly to use gaze cues, as reviewed in the next section. In light of the need to cope with sensorimotor delays [10], gaze control itself in task-based scenarios can be considered as inherently predictive of a number of action-relevant aspects. Indeed, in moving our eyes we make use of knowledge- and sensorimotor-based experience ([11], [12], [13], [14]) to quickly retrieve the information needed to plan the next limb motion.

In this preliminary study we focus on gaze-based intention prediction in teleoperating a robotic gripper in a simulated scenario, in order to investigate human eye-hand coordination under these conditions and devise an intention recognition model to be later transferred to a real world shared autonomy scenario. As a first setup for object manipulation we concentrate on basic pick-and-place tasks as common in this kind of architectures ([15], [5], [16], [17], [18]). Presented contributions are a behavioral assessment of eye-hand coordination in such scenarios and an instantiation of a Hidden Markov Model trained on collected data, showing good generalizability across users and task configurations.

In the next sections, related work on gaze-based intention recognition is first reviewed; the experimental methods used in our setup and the devised models are further presented, followed by results obtained from behavioral analysis and model testing. We conclude discussing emerged implications and future perspectives.

[1] Stefan Fuchs and Anna Belardinelli are with the Honda Research Institute Europe, Offenbach, Germany stefan.fuchs,anna.belardinelli@honda-ri.de

II. RELATED WORK

That the task shapes the way we look at the world has long been known, as shown in [19]. In that study, it was shown that depending on the question the viewer was trying to answer different scanning patterns were produced on the very same image. A number of studies have replicated and confirmed Yarbus' experiment and managed to invert the process and estimate the task from eye data above chance level (e.g., [20], [21], [22]). The most popular and effective techniques to compute the probability of a given task given eye movements and possibly their sequence entail Naive Bayes classifier, Hidden Markov Models, SVM, multivariate pattern analysis and random forests (see for a more complete review [23]). The largely increased diffusion of wearable cameras and eye-trackers in recent years has triggered research on daily activities recognition as observed from an egocentric perspective [24], [25], [26], hence relying on eye, hand, head and possibly body coordination (see [27] for a full review). Yet the approaches above are concerned either with passive information-seeking or with activity recognition rather than with simple action or intention recognition. In human-robot collaboration often the robot partner is aware of the activity context and for effective cooperation it just needs to detect the current action intention of the human partner in order to help them with it. Indeed, there are two basic types of intention [28]: a mental state of an intention for the future (distal intention) and intentionality for an immediate action (proximal intention). From a temporal perspective a proximal intention is very close to the executed action. Thus, the boundary between intention recognition and action recognition is very blurry. The later an intention is recognized the more advanced might be the execution of an action.

Huang and Mutlu [29] have proposed a method for anticipatory control which allows a robot to predict the intent of the human user and plan ahead of the explicit command. In the task considered, a robotic arm prepares a smoothie by picking the ingredients selected vocally by a human user looking at an illustrated list. By means of eye tracking the robot infers the user intention before they utter it and anticipates picking the intended ingredient: an SVM was fed a feature vector of gaze features for each ingredient, such as the number of glances, duration of the first glance, total duration and whether it was the most recently glanced item as predictors of the currently intended ingredient. Although such an approach seems simple and effective in this case the human user was carrying out no parallel visuomotor control task that could yield spurious fixations.

Haji Fathaliyan and colleagues [30] propose a method to localize gaze on 3D objects by projecting the gaze vector on point cloud rep- resentations of objects manipulated by a person preparing a powdered drink. The authors produced then 3D heat maps displaying the most gazed locations on each object depending on the performed subtask. By means of Dynamic Time Warping barycentric averaging, sequences of gazed objects were obtained encapsulating the typical temporal patterns of object interaction that could be used for action recognition.

Within shared autonomy approaches, as a first attempt at integrating gaze input from the user, Admoni and Srinivasa [31] put forward a proposal relying on Javdani's framework [5], where the probability distribution over the goals (hidden states) is updated by considering both user's eye movements and joystick commands as observations in a POMDP, using hindsight optimization to solve it in real time.

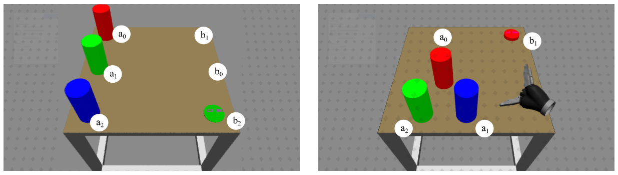
In a further study [32], the authors present a preliminary eye-tracking experiment aimed at comparing user behavior within-subjects in different teleoperation modalities, namely with more or less autonomy. In the scenario of an assistive robot arm spearing food bits from a plate to feed an impaired user, by looking at partly manually annotated gaze behavior, two patterns of fixations emerged: monitoring glances, meant to check the translational behavior of the arm approaching the intended food morsel, and planning glances, which select the target morsel before starting the arm actuation, as in natural eye-hand coordination ([33], [34]). In the context of assistive robotics, a number of studies have also considered gaze information (also combined with multimodal interfaces such as BCI and haptic feedback) to operate robotic limbs and wheelchairs ([35], [36], [37])

To investigate human oculomotor behavior during teleoperation in a more controlled scenario and with a more natural input interface, we designed an experiment in simulation, where the participant would control the remote robot arm by means of their movement via motion tracking. We reasoned that this would produce more natural scanpaths and reaching behavior, without the cognitive overload of a controller with few DOFs. These behavioral cues were collected to train a proof-of-concept model able to predict the current intention in pick-and-place tasks, to be later deployed in a real-world setup.[1] Since many teleoperation scenarios relay visual input through a camera, we displayed the scene on a screen and used eye tracking glasses to retrieve the point-of-regard (POR) on the 2D display.

In a very recent study [38] considering object aligning tasks in Virtual Reality, it was shown how already simple features such as the proportion of PORs on distinct Areas-of-Interest (AoIs) within the objects could constitute a sufficient oculomotor signature to discriminate between four different tasks, which could be classified well above chance.

In our approach, since we plan to work with multiple objects and to recognize different sequential sub-tasks, we chose to model scanpaths via Hidden Markov Models (HMM), which present the benefit of considering the temporal dimension of the gaze shifts and can better deal with spurious fixations and gaze samples and varying eye tracking frequency ([39], [40], [41]).

[1] To avoid confusion with terms sometimes used interchangeably, sometimes meaning different things, we here refer to task as the overarching ongoing activity, e.g. pick and place, while intention implies the commitment to perform the current proximal action/sub-task, e.g. reaching to grasp.



(a) Lined-up Arrangement (b) Triangle-Shape Arrangement

Fig. 2: Example of a scene used in each trial. The objects to pick up were displayed on the left side in 3 different colors while the disc on the right could similarly appear at each of 3 positions on the right side. The color of the disk signified which cylinder was to pick up, the position of the disk denoted the position for the placing down.

III. EXPERIMENTAL METHODS

A. Participants

This study has been conducted after the outbreak of the CoViD-19 pandemic. Hence, thus far, a number of participants suitable for this kind of study could not be recruited and to minimize infection risks only associates of the Honda Research Institute participating in this project were asked to take part in data collection on a voluntary basis (N = 4, including the authors). We complied with the measures of the *Occupational Safety and Health Standard* emanated by the German Federal Ministry of Labour and Social Affairs by keeping a safe distance and wearing face masks. The study was approved by the Bioethics Committee of Honda.

Participants had normal or corrected-to-normal vision, were all right-handed and gave informed consent to participating in the study.

B. Experimental setup and procedure

The experiment was carried out in a simulated scenario created with the Gazebo Simulator (see Fig. 1). The scene was captured with a virtual sensor and displayed on a wide screen (1.21 m × 0.68 m) with HD resolution in front of the participant, who was standing at a distance of about 1.5 m. Participants wore a binocular Pupil Core eye-tracker by Pupil Labs, working at 100 Hz with a reported accuracy of 0.6 °. They also held in the right hand the HTC Vive controller, tracked by the Vive Lighthouse system for input control in the teleoperation task. All physical devices and surfaces were sanitized after each use.

After instructions, participants were required to wear the eye tracking glasses, to adjust the eye and scene cameras according to the experimenter directions and to perform a 5-points calibration.

The experimental stimuli consisted of 3 cylinders presented in two configurations (in different blocks): either aligned on the left side of a table (numbered as follows: 0 for the top, 1 for the middle, 2 for the bottom of the table) or at the vertices of a virtual triangle (0 for the top vertex, 1 for the bottom right, 2 for the bottom left; see Fig. 2). Colors were permuted anew in each trial. Along with the cylinders a disk would appear on the right side of the table, at one of three positions (denoted as: 0 top, 1 middle, 2 bottom). The disk specified the current pick-and-place task: the color indicated which cylinder to pick up and the position of the disc where the cylinder had to be placed down on the table. The task would be executed by a robotic gripper in the virtual scene, operated by the participant movements. Participants were required to reach with the controller in their hand toward the target and to pick it up by pressing the button on the controller under the index finger. They had then to move the cylinder to the other side and release it on the place position, in so ending the trial. Between trials a resting time of 5 s was given, followed by a fixation cross and indications on how to move the controller back to the rest position. As soon as the controller reached the starting position, the next trial started.

C. Design and data processing

We designed two different arrangements of the cylinders: First, a basic arrangement features three cylinders lined up on the left hand side of the table (see Fig. 2a). Second, the cylinders are arranged in a triangle. Thus, we can investigate the impact of the spatial arrangement of the items on the gaze behavior.

In each trial the target pick and place positions are randomly generated. Instead of working with relative eye coordinates, we used the fiducial markers and the scene camera of the Pupil Labs device to localize the eye-tracking-glasses in the scene w.r.t. the world and screen, respectively. Fixations represent a very popular cue in eye-movement data analysis and might seem an obvious choice in this intention estimation application. The parameterization of a fixation identification method, however, might be very arbitrary. Usually, it is not clear and agreed on, when exactly fixations start and when they end. Thus, the parameters of a fixation identification algorithm can have a dramatic impact on our higher-level analyses [42]. Further, the system will be required to work online eventually and online fixation recognition is not always accurate while further increasing the computational load. The temporal information related to dwelling time in the AoIs (the objects of interest in the scene) during fixations is still learnt and considered by the HMM all along.

For these reasons, we decided to work with gaze samples that were mapped on the scene according to the following approach: A heatmap with a discrete resolution represents the hemispherical field-of-view of the participant. In this case a sampling of 1 ° is used and the heatmap comes with a resolution of 180 px by 90 px. The user's eye gaze $\mathbf{g}$ is represented by a two-dimensional normal distribution and the density is plotted onto the heatmap with gaze uncertainty σ and direction μ.[2] As an example: A gaze collinear with head orientation comes with a density maximum plotted in the center of the heatmap. The size of σ depends on the accuracy and precision of the eye tracking measurements.

[2] The gaze was mapped in this way since in a later stage we plan to move the simulation into a virtual reality headset with embedded eyetracking and the gaze mapping on the scene can stay unaltered.

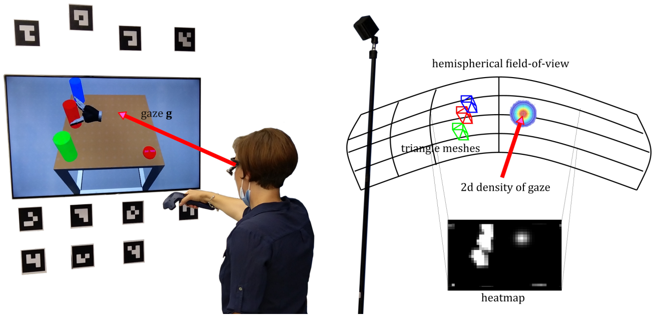


Fig. 3: The user's field-of-view is approximated by a hemispherical heatmap. The density of a 2d normal distribution centered on the point-of-regard represents the gaze and its uncertainty. The surface integral over the triangles of a certain object is the likelihood of this AoI.

We set $\sigma = 2^\circ$ which is in accordance with the size of the human fovea. All potential scene objects are represented as triangle meshes with a bounding box made of at least 12 triangles. The pose of the objects is known from a scene understanding module and given the localization of the eye-tracking glasses the object poses can be transformed into the head coordinate system. Triangles, that are visible to the user (i.e. normal of triangle directed towards user), are plotted into the heatmap. The surface integral of the density function represents the likelihood that this area has been regarded by the user. The complete likelihood (of each object to be regarded by the user) is the sum of all visible triangles the object is made of. In order to not overemphasize large objects, all likelihoods are normalized by their visible areas. For each object an Area-of-Interest was defined, for a total of 7 AoIs: for the picking objects the areas $\{a_0, a_1, a_2\}$, for the placing positions the areas $\{b_0, b_1, b_2\}$, plus an area R for the robotic gripper.

As a result, this so-called *Area-of-Interest-analysis* provides for every gaze sample $\mathbf{g}$ a feature vector $\mathbf{F}$ entailing the likelihood computed for each of these AoIs:

$$\mathbf{F}_t = \{\mathrm{P}(\mathrm{AoI} = a_0|\mathbf{g}_t), \mathrm{P}(\mathrm{AoI} = a_1|\mathbf{g}_t), ...\} \quad . \qquad (1)$$

These were logged along with the current hand position and robot gripper position and with the current grasping state (defined as the binary state of the grasping button). Moreover, each trial was labeled with a Boolean feature to state if it was successful. Indeed, if the grasp failed for any reason multiple grasp attempts could be observed or none at all if the cylinder was toppled down and fell off the table.

D. Modeling Intentions with a Gaussian HMM

Our approach aims at predicting the proximal intention, i.e. the current action and the involved object.

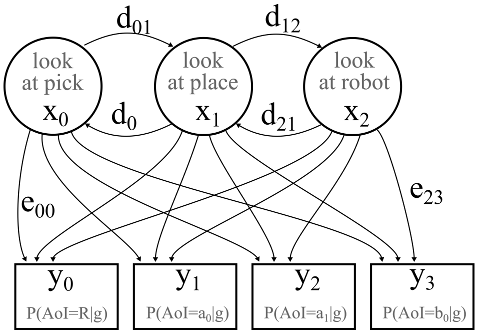


Fig. 4: Exemplary Gaussian Hidden Markov Model for a Pick-and-Place-Task. d refers to the state transition probabilities and the hidden states X might be looking at the object to be picked, at the robot or at the placing position target. e refers to the emission probabilities for the possible observations Y, e.g. looking at the robot, the objects in the scene $\{a_0, a_1\}$ or the release goal position b_0.

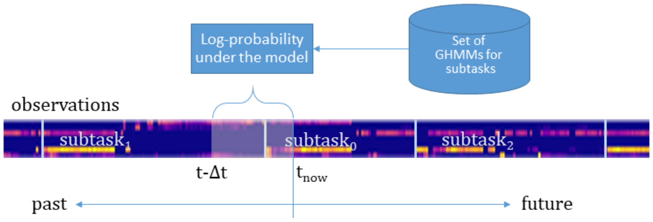


Fig. 5: The observations made in the last Δt seconds are used to compute the log-probability of these observations under each of the trained GHMMs. The log-probabilities (presented on the vertical axis of the plot) are an indication for the respective intention (lighter color represent higher probability). The length Δt of the time window decides on the accuracy and the earliness of the intention predictions.

Gaze not only comes with a specific pattern during action execution but also provides early cues that indicate parameters of a pick and place task, such as which object to pick or where to place it down. These parameters are defined by the proximal intention. The temporal gaze pattern can be represented with a Hidden Markov Model (see Fig. 4). The hidden states $\mathbf{X}(t)$ describe the internal intention process and might relate to *looking at the target object* or *looking at the placing position*. However, this is just an assumption, while the hidden Markov process drives an observable gaze sequence $\mathbf{Y}(t)$. The gaze sequence is described by the AoI likelihoods as derived from the multivariate Gaussian distribution (see Sec. III-C). The distribution of these AoI likelihoods at a particular time is governed by the emission probabilities of the hidden Markov process given the state of the hidden variable at that time. This approach is independent of the gaze sequence length, i.e. observation sampling and execution velocity, as long as the sequences are scaled linearly.

We defined 6 intentions to be recognized: 3 pick-up intentions (for each of the 3 cylinders) and 3 place intentions (for each of the 3 placing positions). Hence, 6 HMMs have been configured with 5 internal states. The observation vector of a HMM comprises 8 components: the AoI likelihoods of the 3 cylinders, the AoI likelihoods of the 3 possible placing positions, the AoI likelihood of the robot, and the trigger button state of the Vive controller. The transition and emission parameters were learned for each HMM given between 19 and 31 observations sequences (for a total of 160) for the respective actions collected from 2 users. The training is done offline with data only from the lined-up arrangement and successful pick-and-place tasks.

Fig. 5 sketches the online intention recognition approach. At every time step t the observations from the last Δt seconds are used to compute the log-probability of these observations under each of the trained HMMs. The HMM with best log-probability exceeding a given threshold ($\kappa > 0$) is taken as prediction of the respective intention. If no model scores over the threshold, no intention is confidently recognized. The offline training and the online recognition are implemented in Python with the help of the *hmmlearn*-library[3].

The performance of this approach is tested on data from 4 users and between 17 and 28 observation sequences for each intention, respectively (for a total of 128 sequences). The testing data comprised unseen sequences from the two users used for training plus sequences from two additional users. Moreover, testing was done also on sequences from blocks with triangular arrangement (between 19 and 33 sequences for each intention, for a total of 156).

IV. RESULTS

A. Behavioral analysis

To get a better picture of the gaze behavior during the presented task, we looked at some behavioral measures, seeking confirmation of some of the patterns described in [32]. Due to the low number of participants thus far, we could not perform an inferential statistics analysis to test any hypothesis, hence we report a descriptive analysis computed over the whole dataset depending on the different tasks.

Two exemplary trajectories for different pick-and- place tasks and object configurations are depicted in Fig. 6. At any time, the AoI collecting the highest likelihood is considered the one currently looked at. It can be noted that upon motion onset the AoI corresponding to the place target (whose color determines also the picking target) is glanced, that is, this is a planning glance, as defined in [32], while afterwards the gaze moves to the picking target. After picking, the gaze moves to the placing target. During both the reaching to grasp phase and the transport to place phase the robot AoIs (gray) is checked in a monitoring pattern, to make sure the gripper is moving in the intended direction.

For each trial we consider two intentions/(sub)tasks, one picking and one placing intention, separated by the key press triggering the grasping. To get a more complete overview of the time the gaze spent in different AOIs across tasks, the relative time distribution of gaze on each AoI was computed and is presented in Fig. 7. To make the picture easier to interprete, we considered that for each intention there are actually just 5 semantic entities that can used to describe the AOIs, namely: the pick target (e.g.a_0 for *pick_0*), the pick distractors (e.g.a_1, a_2 for *pick_0*), the place target (e.g. any of the b_i AOIs depending on the current task), the place distractors (e.g. any of the b_i AOIs that are not the target). Analogously for the place intention, the place target would be the specific AOI related to that task, while the pick target could be any of the picking positions and the distractors are the pick and place AOIs that are not target of the trial.

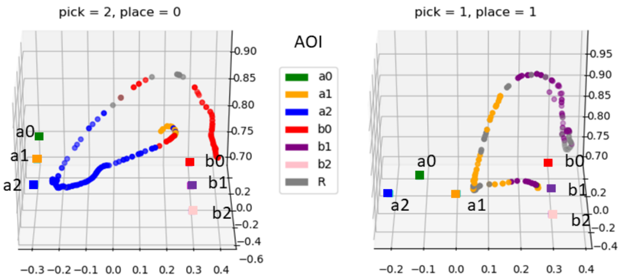


Fig. 6: Two exemplary trajectories of the hand during the pick-and-place tasks (left: pick in position 2 and place in position 0; right: pick in position 1 and place in position 1). The movement samples are colored with the currently gazed AoI (see legend). The square markers denote the picking and placing positions, for clarity here denominated and colored as the respective AOIs.

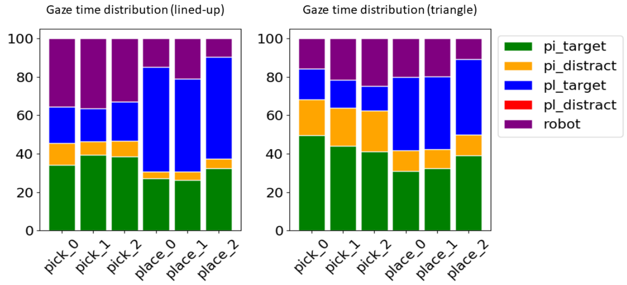


Fig. 7: Relative distribution of the time the gaze spent on semantic AoIs across tasks for the lined-up (left) and the triangle arrangement (right). In the pick tasks the respective picking AoIs (pi_target) are more looked at, in the place tasks the respective placing AoIs (pl_target).

As it can be noted, each task presents a relatively similar distribution on a semantic level, with most of the time spent on the current pick or place target, yet the distributions are quite distinctive considering that each action target is the AoIs related to the task (that is the corresponding a_i position in the pick tasks and the corresponding b_i position in the place tasks). In the pick tasks the the place target is briefly looked at to learn the picking task, while in the place tasks the pick target receive also some attention since, after pressing the button for the grasp and in absence of any haptic feedback, the gaze checks that the object is successfully moving along with the gripper. Interestingly, in both configurations and tasks also the robot effector receives a discrete amount of gaze time. In normal eye-hand coordination the hand instead is rarely looked at because proprioceptive information and peripheral vision usually suffice to monitor it. This suggests that in this teleoperation scenario the unusual sensorimotor

[3] https://hmmlearn.readthedocs.io

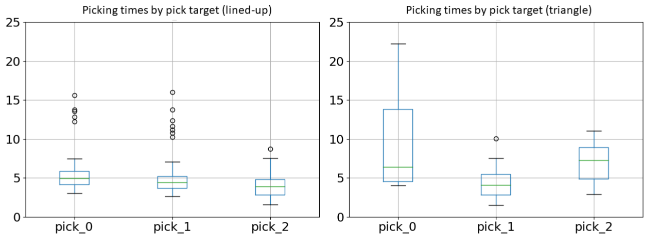


Fig. 8: Picking times across picking position for the two object configurations. The rearmost positions requires a longer reaching time in both configurations, also due to difficult depth estimation. In the triangle configuration the forefront position on the left (pick_2) requires a longer reaching

mapping from the arm and controller to the three-fingered robotic gripper, especially considering the grasp pose, and possibly some delay in the tracking make the user uncertain about the effector movements and current pose. Participants, thus, produced multiple monitoring glances [32] during the movements to visually adjust the effector trajectory and pose. However, in general the distributions looked rather distinctive across tasks, suggesting that it could be possible to reliably discriminate among them, while they looked rather similar across picking configurations, hinting to the possibility to generalize from one to the other. The pick distractors are looked at especially during picking, since the gaze checks the neighboring cylinders in order to decide the best grasp and in order not to collide with them. This is especially the case in the triangle configuration since the cylinders are all close to one another. The place distractors do not receive any attention since in each task only the target position was made visible with a disk (see Fig. 2).

To gain further insight into the difficulty of the task, we looked into the number of failed trials across picking tasks. Error rates were computed for the three pick tasks in the two configurations. The picking action in the lined-up (triangle) configuration was successful in the 71.4% (68.6%) of *pick_0* cases, 88.9% (88.6%) for *pick_1* trials, and 79.6% (85.2%) for *pick_2*. That is, the users could accomplish the task in the vast majority of the cases, still a significant number of failed grasps occurred when picking at position 0 in both configurations.

This could be the case for different reasons: in the lined-up configuration the 0 position is the rearmost and the one requiring to stretch the arm until the furthest edge of the table, yet 3D depth on a 2D plane is badly estimated, especially in the virtual scene where size cues are more difficult to gauge and the own body could not either be used as reference; in the triangle configuration the 0 position is closer to the user yet the other 2 blocks are placed in front of it, requiring to pick the cylinder from above or -for a right-handed user- trying to avoid the cylinder in position 1 going around it. The depth estimation difficulty could yet be ameliorated in a virtual reality set-up.

A similar pattern emerges also looking at picking times (considered as the time from start of the trial to the grasp detected via button press). In this case we considered only successful trials, since in a failed trial no grasp or more than one grasp could occur. Looking at Fig. 8, it can be noted again that the rearmost position requires the longest reaching time. In the case of the triangle configuration, also items in position 2 require a more careful movement, since a right-handed person needs to mind avoiding the cylinder in position 1 when approaching the cylinder in position 2 with the open gripper.

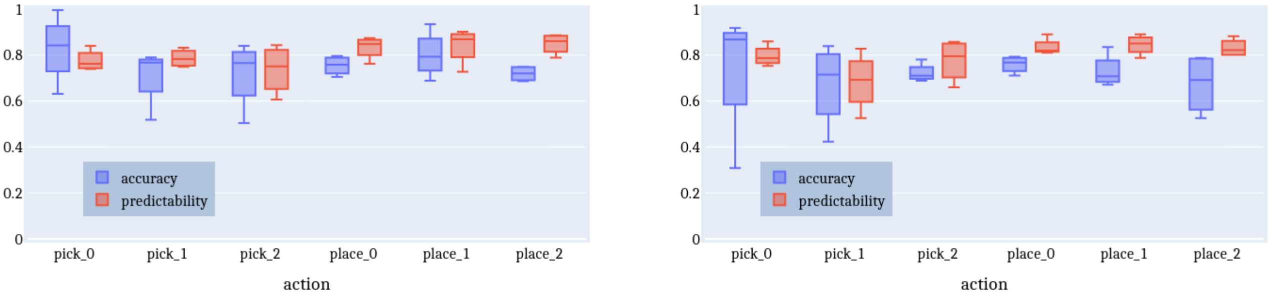
B. Intention recognition

Fig. 9a shows the accuracy and predictability of the intention recognition when using a time window of 0.9 s for the lined-up arrangement. On average, the HMM with the best log-probability being above the given threshold ($\kappa > 0$) indicates the correct intention in 76 % of cases (chance level = 16.7%). Predictability refers to the fraction of action execution time where an intention is recognized (regardless of whether right or wrong). Fig. 10a highlights the relationship between time window Δt, accuracy and predictability. With a longer time window both the prediction accuracy and the predictability decrease. A longer time window has the effect to include more observation samples belonging to a previous action rather than the current intention. This is sketched exemplarily in Fig. 5. As a result, either the log-probability threshold is not exceeded or an incorrect intention is recognized. There is a maximum accuracy at a time window of 0.9 s with a predictability of 80 %. That is, after at least 20 % of the action execution the right action is predicted in 76 % of cases. Given this earliness we can speak of intention recognition.

Fig. 10b plots a similar relationship between time window and performance for the triangle-shaped arrangement. The optimal time window size here is 1.3 s with an accuracy of 72 % (chance level = 16.7%) and a predictability of 79 %. The accuracy curve seems to be flattened, because the action execution times in this setup come with a larger spread. Especially, picking up the cylinders at positions 0 and 1 is more challenging and causes a longer execution time compared to the other sub-tasks in this triangle setup. This issue is apparent also in Fig. 9b with more distant whiskers and extended boxes for *pick_0* and *pick_1*.

Furthermore, the plots in Fig. 10 and Fig. 9 confirm the observations described in Sec. IV-A. The gaze behavior seems to be independent of the spatial arrangement of the objects in the scene. This fact is very well represented by the HMMs, which have been trained only on lined-up arrangement data, but perform almost as well on the triangle arrangement data.

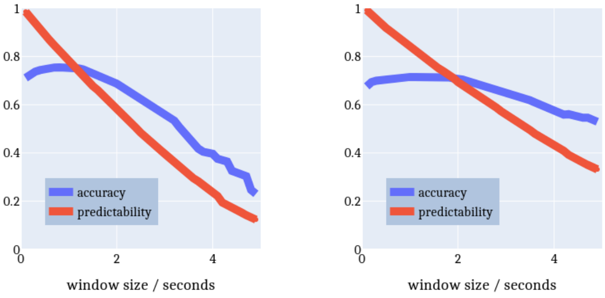
Finally, Fig. 11 shows the confusion matrices for the two tested spatial arrangements. It can be appreciated that when the model delivers a wrong prediction it usually mistakes neighboring picking or placing locations, but still identifies the correct task.



(a) Intention recognition with Δt=0.9s (lined-up)

(b) Intention recognition with Δt=1.3s (triangle shape)

Fig. 9: Accuracy and predictability of online intention recognition for lined-up arrangement and triangle shape arrangement.



(a) lined-up arrangement

(b) triangle arrangement

Fig. 10: Relationship between time window (min = 0.1 s, max = 5 s) and performance measures. Time window and performance are inversely proportional to each other. There is, however a maximum accuracy at time window Δt=0.9 s for the lined-up arrangement and Δt=1.3 s for the triangle arrangement, respectively.

V. DISCUSSION AND CONCLUSIONS

We presented a study aimed at investigating eye-hand coordination and gaze-based intention recognition during teleoperated pick-and-place tasks. The ultimate goal is to transfer such intention recognition into a shared autonomy architecture. To this end, in this first study data was collected, analyzed and modeled in order to have on the one hand a baseline characterization of user behavior in a fully teleoperated modality, on the other hand to train a model flexible enough to work with different users and in possibly different settings.

The analysis of eye and hand behavior has revealed that, although participants managed to successfully operate the gripper in the pick-and-place task in most cases, still some positions required more grasp attempts and longer reaching times. This in part due to the impaired depth estimation on the screen, however the difficulty in aligning the gripper with the cylinder in the furthest position or in avoiding bumping into cylinder 1 to grasp in position 2 in the triangular arrangement required extra care and slowed down the movement.

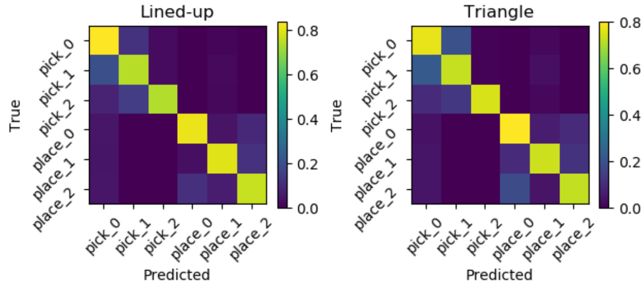


Fig. 11: Normalized confusion matrices for the two picking arrangements. Error are made mostly mistaking neighboring locations but still classifying the task correctly.

Moreover, while the gaze behavior showed some similarities with natural eye-hand coordination, e.g. locating and guiding the hand to the target of the next proximal intention [43], we found that both in the reaching and in the transport phase the robot gripper was looked at for quite some time, differently from what happens when grasping with the own hand. This represents an indicator that the participant preferred to visually monitor the gripper movement in the absence of the usual proprioceptive coordination and tactile feedback. Furthermore, the object held in hand was looked at also after the grasping was triggered, again something that does not happen in natural eye-hand coordination, since tactile feedback confirms the expected contact event and successful grasping [44]. In this teleoperation scenario instead the grasp had to be confirmed visually, hence the gaze lingered on the picked object and only after seeing the object moving along with the hand it moved on to the next distal intention (i.e. the placing position). These effects will be hopefully confirmed as significant once more participants are tested. Still, these kinds of measures offer an insight into the user experience of the teleoperation task: as long as the uncertainty about the task execution is high, the gaze is less anticipative and lingers there where further information needs to be acquired to carry out the task. Although some of these issues could be mitigated with longer training, allowing the user to master the new visuomotor mapping and task [45], an intention recognition model embedded in a shared autonomy architecture that could adjust the robot movement and grasping pose to reliably produce the intended grasp would shorten these training times, allow a more natural eye-hand coordination and relieve the gaze system of monitoring every sub-task unfolding and transition with extra care. That is, an effective shared autonomy system would be validated by shorter execution times, less failed grasp attempts and more anticipative gaze behavior with less time spent monitoring the grasped object and the robot gripper. This would confirm that the user trusts the robotic partner to correctly infer and assist with the current intention but that their sense of agency is preserved, since they anticipate the next subtask in their plan.

Apart from these considerations, as shown for a different task [38], we also found that the gaze behavior was reliably different across tasks and could be hence learnt and predicted effectively. To this end a Hidden Markov Model was devised for each of the intentions to be recognized, considering as emissions the normalized likelihood of the gaze (represented as a Gaussian distribution) to be on each of the objects in the scene, including the robot hand. The system was trained on pick-and-place tasks from 2 users and then tested on similar sequences from the 2 users plus 2 other users. Considering a time window of 0.9 s where emissions are accumulated and then scored by the 6 HMMs, the model achieves a well above chance accuracy across all tasks, returning a prediction as early as after seeing 20% of the current action, on average. Beside with different users, the generalizability of the system was further tested on a different geometrical configuration of the pick task, delivering comparable accuracy and predictability. This suggests, as already the similar semantic distributions of gaze time across tasks and configurations, that there is a certain invariancy in the gaze patterns, which are mainly shaped by the task at a higher level. That is, at least in simple manipulation tasks and object configurations, sequences of gaze glances at objects are more heavily determined and constrained by the current subtask structure, once the target is specified, rather than by the contingent physical setup. That is, also the oculomotor plan subserving and directing the motor plan seems to reflect the syntactic structure of action [46].

These are promising results for the further development of our intention recognition system and for its embedding in a real-world shared autonomy scenario. Current and future work is going to expand both the training and testing with multiple participants as well as considering more and different objects and tasks.

References

[1] M. A. Goodrich, J. W. Crandall, and E. Barakova, "Teleoperation and beyond for assistive humanoid robots," *Reviews of Human factors and ergonomics*, vol. 9, no. 1, pp. 175–226, 2013.

[2] J. M. Beer, A. D. Fisk, and W. A. Rogers, "Toward a framework for levels of robot autonomy in human-robot interaction," *Journal of human-robot interaction*, vol. 3, no. 2, pp. 74–99, 2014.

[3] M. Schilling, S. Kopp, S. Wachsmuth, B. Wrede, H. Ritter, T. Brox, B. Nebel, and W. Burgard, "Towards a multidimensional perspective on shared autonomy," in *2016 AAAI Fall Symposium Series*, 2016.

[4] W. Yu, R. Alqasemi, R. Dubey, and N. Pernalete, "Telemanipulation assistance based on motion intention recognition," in *Proceedings of the 2005 IEEE international conference on robotics and automation*. IEEE, 2005, pp. 1121–1126.

[5] S. Javdani, S. S. Srinivasa, and J. A. Bagnell, "Shared autonomy via hindsight optimization," in *Robotics science and systems: online proceedings*, vol. 2015. NIH Public Access, 2015.

[6] K. Hauser, "Recognition, prediction, and planning for assisted teleoperation of freeform tasks," *Autonomous Robots*, vol. 35, no. 4, pp. 241–254, 2013.

[7] D. Aarno and D. Kragic, "Motion intention recognition in robot assisted applications," *Robotics and Autonomous Systems*, vol. 56, no. 8, pp. 692–705, 2008.

[8] A. K. Tanwani and S. Calinon, "A generative model for intention recognition and manipulation assistance in teleoperation," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 43–50.

[9] T. Yang, W. Huang, C. K. Chui, Z. Jiang, and L. Jiang, "Stacked hidden markov model for motion intention recognition," in *2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP)*, 2017, pp. 266–271.

[10] R. Miall and G. Reckess, "The cerebellum and the timing of coordinated eye and hand tracking," *Brain and cognition*, vol. 48, no. 1, pp. 212–226, 2002.

[11] J. M. Henderson, "Gaze control as prediction," *Trends in cognitive sciences*, vol. 21, no. 1, pp. 15–23, 2017.

[12] M. M. Hayhoe, "Vision and action," *Annual review of vision science*, vol. 3, pp. 389–413, 2017.

[13] K. Fiehler, E. Brenner, and M. Spering, "Prediction in goal-directed action," *Journal of vision*, vol. 19, no. 9, pp. 10–10, 2019.

[14] A. Belardinelli, M. Barabas, M. Himmelbach, and M. V. Butz, "Anticipatory eye fixations reveal tool knowledge for tool interaction," *Experimental brain research*, vol. 234, no. 8, pp. 2415–2431, 2016.

[15] S. Li, X. Zhang, and J. D. Webb, "3-d-gaze-based robotic grasping through mimicking human visuomotor function for people with motion impairments," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 12, pp. 2824–2835, 2017.

[16] A. Shafti, P. Orlov, and A. A. Faisal, "Gaze-based, context-aware robotic system for assisted reaching and grasping," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 863–869.

[17] S. Jain and B. Argall, "Recursive bayesian human intent recognition in shared-control robotics," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 3905–3912.

[18] ——, "Probabilistic human intent recognition for shared autonomy in assistive robotics," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 9, no. 1, pp. 1–23, 2019.

[19] A. L. Yarbus, "Eye movements during perception of complex objects," in *Eye movements and vision*. Springer, Boston, MA, 1967, pp. 171–211.

[20] A. Borji and L. Itti, "Defending yarbus: Eye movements reveal observers' task," *Journal of vision*, vol. 14, no. 3, pp. 29–29, 2014.

[21] A. Haji-Abolhassani and J. J. Clark, "An inverse yarbus process: Predicting observersâ^{TM} task from eye movement patterns," *Vision research*, vol. 103, pp. 127–142, 2014.

[22] C. Kanan, N. A. Ray, D. N. Bseiso, J. H. Hsiao, and G. W. Cottrell, "Predicting an observer's task using multi-fixation pattern analysis," in *Proceedings of the symposium on eye tracking research and applications*, 2014, pp. 287–290.

[23] J. F. Boisvert and N. D. Bruce, "Predicting task from eye movements: On the importance of spatial distribution, dynamics, and image features," *Neurocomputing*, vol. 207, pp. 653–668, 2016.

[24] W. Yi and D. Ballard, "Recognizing behavior in hand-eye coordination patterns," *International Journal of Humanoid Robotics*, vol. 6, no. 03, pp. 337–359, 2009.

[25] A. Fathi, Y. Li, and J. M. Rehg, "Learning to recognize daily actions using gaze," in *European Conference on Computer Vision*. Springer, 2012, pp. 314–327.

[26] K. Ogaki, K. M. Kitani, Y. Sugano, and Y. Sato, "Coupling eye-motion and ego-motion features for first-person activity recognition," in *2012*

IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. IEEE, 2012, pp. 1–7.

[27] T.-H.-C. Nguyen, J.-C. Nebel, F. Florez-Revuelta *et al.*, "Recognition of activities of daily living with egocentric vision: A review," *Sensors*, vol. 16, no. 1, p. 72, 2016.

[28] M. Bratman *et al.*, *Intention, plans, and practical reason*. Harvard University Press Cambridge, MA, 1987, vol. 10.

[29] C.-M. Huang and B. Mutlu, "Anticipatory robot control for efficient human-robot collaboration," in *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*. IEEE, 2016, pp. 83–90.

[30] A. Haji Fathaliyan, X. Wang, and V. J. Santos, "Exploiting three-dimensional gaze tracking for action recognition during bimanual manipulation to enhance human–robot collaboration," *Frontiers in Robotics and AI*, vol. 5, p. 25, 2018.

[31] H. Admoni and S. Srinivasa, "Predicting user intent through eye gaze for shared autonomy," in *2016 AAAI Fall Symposium Series*, 2016.

[32] R. M. Aronson, T. Santini, T. C. Kübler, E. Kasneci, S. Srinivasa, and H. Admoni, "Eye-hand behavior in human-robot shared manipulation," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018, pp. 4–13.

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

[34] M. M. Hayhoe, A. Shrivastava, R. Mruczek, and J. B. Pelz, "Visual memory and motor planning in a natural task," *Journal of vision*, vol. 3, no. 1, pp. 6–6, 2003.

[35] H. Zeng, Y. Shen, X. Hu, A. Song, B. Xu, H. Li, Y. Wang, and P. Wen, "Semi-autonomous robotic arm reaching with hybrid gaze–brain machine interface," *Frontiers in Neurorobotics*, vol. 13, p. 111, 2020.

[36] M. Wang, A. A. Kogkas, A. Darzi, and G. P. Mylonas, "Free-view, 3d gaze-guided, assistive robotic system for activities of daily living," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 2355–2361.

[37] V. Schettino and Y. Demiris, "Inference of user-intention in remote robot wheelchair assistance using multimodal interfaces," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 4600–4606.

[38] A. Keshava, A. Aumeistere, K. Izdebski, and P. Konig, "Decoding task from oculomotor behavior in virtual reality," in *Symposium on Eye Tracking Research and Applications*, 2020, pp. 1–5.

[39] A. Belardinelli, F. Pirri, and A. Carbone, "Bottom-up gaze shifts and fixations learning by imitation," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 37, no. 2, pp. 256–271, 2007.

[40] G. Boccignone, "Advanced statistical methods for eye movement analysis and modelling: a gentle introduction," in *Eye Movement Research*. Springer, 2019, pp. 309–405.

[41] A. Coutrot, J. H. Hsiao, and A. B. Chan, "Scanpath modeling and classification with hidden markov models," *Behavior research methods*, vol. 50, no. 1, pp. 362–379, 2018.

[42] D. Salvucci and J. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," 01 2000, pp. 71–78.

[43] 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.

[44] R. S. Johansson and J. R. Flanagan, "Coding and use of tactile signals from the fingertips in object manipulation tasks," *Nature Reviews Neuroscience*, vol. 10, no. 5, pp. 345–359, 2009.

[45] U. Sailer, J. R. Flanagan, and R. S. Johansson, "Eye–hand coordination during learning of a novel visuomotor task," *Journal of Neuroscience*, vol. 25, no. 39, pp. 8833–8842, 2005.

[46] K. Pastra and Y. Aloimonos, "The minimalist grammar of action," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 367, no. 1585, pp. 103–117, 2012.