# Rebellion and Obedience: The Effects of Intention Prediction in Cooperative Handheld Robots

Janis Stolzenwald and Walterio W. Mayol-Cuevas

*Abstract*— **Within this work, we explore intention inference for user actions in the context of a handheld robot setup. Handheld robots share the shape and properties of handheld tools while being able to process task information and aid manipulation. Here, we propose an intention prediction model to enhance cooperative task solving. The model derives intention from the combined information about the user's gaze pattern and task knowledge. Within experimental studies, the model is validated through a comparison of user frustration for the case where the robot follows the predicted location of the user's intended action versus doing the opposite (rebellion). The proposed model yields real-time capabilities and reliable accuracy up to** 1.5 s **prior to predicted actions being executed.**

## I. Introduction

A handheld robot shares properties of powered hand tools while being enhanced with autonomous motion as well as the ability to process task-relevant information and user signals. Since the robot holds task knowledge, such a system could help cutting workers' training times, as less user expertise is required for task solving. At the same time, the robot benefits from humans' natural navigation and obstacle avoidance capabilities. While this can arguably be beneficial for the task performance, the high proximity between the user and the robot also leads to codependencies that create the need of communication methods between the user and the robot for efficient collaboration.

Earlier work in this field explored robot-human communication for improved cooperation [1], [2]. Such one-way communication of task planning, however, is limited in that the robot has to lead the user and as users exert their will and decisions, task conflicts emerge. This, in turn, inflicts user frustration and decreases cooperative task performance. We argue, that this is due to a lack of human-robot communication.

As a starting point of addressing this problem, we introduced extended user perception in earlier work on handheld robot collaboration [3]. This allows the robot to estimate the user's point of attention via eye gaze in 3D space during task execution. While the estimation of users' visual attention helps just-in-time planning, we lack an intention model which would allow the robot to infer the user's goal in the proximate future and go beyond reacting to immediate decisions only.

In recent years, promising solutions for intention inference have been achieved through observing user's eye gaze [4], body motion [5] or task objects [6]. These works target safe interactions between humans and sedentary robots with shared workspaces. To our knowledge, they were never tested in a setup of close codependency, such as we face within handheld robotic systems. Hence, this is explored in our research, which is guided by the following research questions:

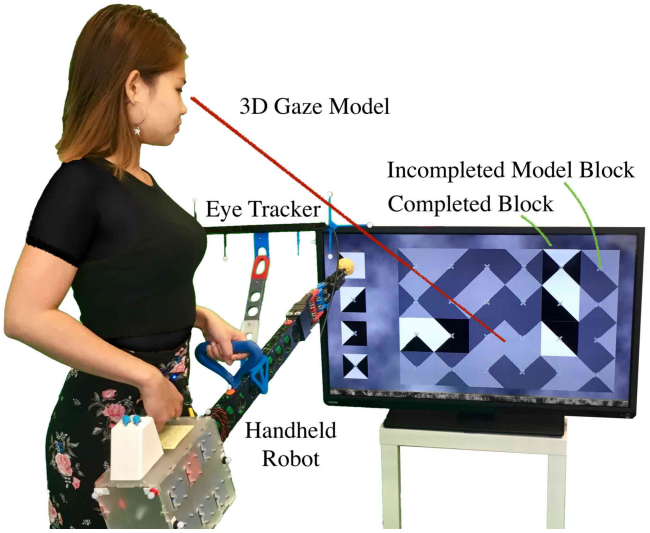



Fig. 1: This picture shows a participant within our user intention prediction study who solves the assembly task and is about to decide where to place the currently held block. Using the eye tracker, the prediction system extracts the user's gaze pattern, which is used for action prediction.

**Q1** How can user intention be modelled in the context of a handheld robot task?

**Q2** To what extent does intention prediction of users affect the cooperation with a handheld robot?

For our study, we use the open robotic platform, introduced in [7], combined with an eye tracking system as reported in [3]. The 3D CAD models of the robot design are available from [8]. Within a simulated assembly task, which was inspired by [9], eye gaze information is used to predict subsequent user actions. Figure 2 shows an overview of our proposed system. The two principal parts of this study consist of modelling user intention in section III-V, followed by its validation through an assistive pick and place task in section VI-VII. Our main contributions are summarised as follows:

- We propose an online intention model, which predicts users' interaction location targets based on eye gaze and task states.

Department of Computer Science, University of Bristol, UK, janis.stolzenwald@bristol.ac.uk, wmayol@cs.bris.ac.uk

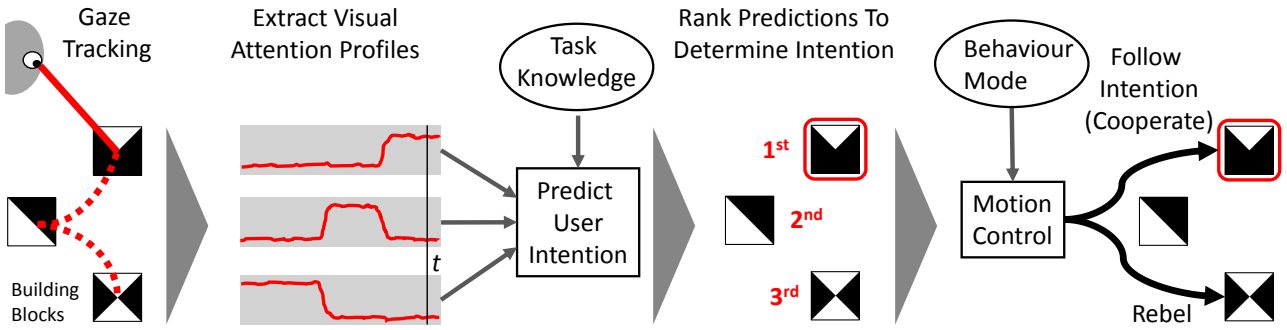



Fig. 2: Overview of the intention prediction model and its use for the robot's motion control.

- For data collection and model validation, we propose an experimental setup of a block copying task to emulate an example of assembly.
- In the absence of universally accepted psychophysical metrics, we propose to measure the frustration induced through the robot's rebellion to validate intention predictions.

## II. Background and Related Work

In this section, we deliver a summary of earlier work on handheld robots and its control based on user perception. Furthermore, we review existing methods for intention inference with a focus on human gaze behaviour.

### A. Handheld Robots

Early handheld robot work [1] uses a trunk-shaped robot with 4-DoF to explore the effect of autonomy on collaborative task performance and perceived task load. This was later upgraded to a 6-DoF (joint space) mechanism [7] and used gestures, such as pointing, to study user guidance. These earlier works demonstrate how users benefit from the robot's quick and accurate movement while the robot profits from the human's tactical motion. Furthermore, it was found that cooperative performance significantly increases when the robot communicates its plans e.g. via a robot-mounted display [2]. However, the work also demonstrates that increased autonomy of the robot can lead to mismatches between user intention and the robot's plan. Sometimes, for example, the robot chose a valid goal, at which time the user decided to move to a different one. This led to irritation and frustration in users on whom the robot's plan was imposed.

Efforts towards involving user perception in the robot's task planning were made in our recent work on estimating user attention [3]. Using a robot-mounted remote eye gaze tracker, the system captures the user's visual attention during task execution. This information is then used to bias the robot's plans towards objects the user focuses on. For tasks with higher speed demands and thus high decision frequencies, the attention-driven mode was rated more cooperative compared to the case where the robot was fully autonomous and was ignoring user attention.

As opposed to an intention model, the attention model would react to the current state of eye gaze information only, rather than using its history to make predictions about the user's future goals. We suggest that intention prediction would be required for cooperative solving of complex tasks like assembly where there is an increased depth of subtasks.

### B. Intention Prediction

Intention estimation in robotics is in part driven by the demand for safe human-robot interaction and efficient cooperation. Ravichandar et al. investigated intention inference based on human body motion [5]. Using Microsoft Kinect motion tracking as an input for a neural network, reaching targets were successfully predicted within an anticipation time of approximately 0.5 s prior to the hand touching the object. The model was later improved using eye gaze tracking for prefiltering, which increased the anticipation time to 0.78 s [10]. Similarly, Saxena et al. introduced a measure of motion-based affordance to make predictions about human actions and reached 84.1%/74.4% accuracy 1 s/3 s in advance, respectively [11].

Huang et al. used gaze information from a head-mounted eye tracker to predict object selections. Using a support vector machine (SVM), an accuracy of approximately 76% was achieved with an average prediction time of 1.8 s prior to the verbal request [12]. In subsequent work, Huang & Mutlu used the model as a basis for a robot's anticipatory behaviour, which led to more efficient collaboration compared to following verbal commands only [4].

We note that, while the above methods improve cooperation in a turn-taking human-robot collaboration setup, we lack knowledge about their effect on cooperation performance within a shared control setup such as we face with handheld robots.

### C. Human Gazing Behaviour

The intention model presented in this paper is mainly driven by eye gaze data. Therefore, we review work on human gaze behaviour to inform the underlying assumptions of our model.

Land et al. found that fixations towards an object often precede a subsequent manual interaction by around 0.6 s [13]. Subsequent work revealed that the latency between eye and hand varies between different tasks [14]. Similarly, Johansson et al. [15] found that objects are most salient for human's when they are relevant for task planning.

The purpose of preceding fixations in manual tasks was furthermore explored through virtual [9] block design tasks. The results show that humans gather information through vision *just in time* rather than memorising e.g. all object locations, which goes in line with work on short-term memory processes [16].

The above work inspired the use of gaze data for action prediction and form the basis of our assumptions for the intention model, formulated in section III-B.

## III. Prediction of User Intention

In this section, we describe how intention prediction is modelled for the context of a handheld robot based on an assembly task.

### A. Data Collection

We chose a simulated version of a block copying task, which has been used in the context of work in hand-eye coordination [9]. Participants of the data collection trials were asked to use the handheld robot (cf. figure 3) to pick blocks from a stock area and place them in the workspace area at one of the associated spaces indicated by a shaded model pattern. The task was simulated on a 40 inch LCD TV display and the robot remained motionless during the data collection task to avoid distraction. We drew inspiration from the block-copy task presented in [9] and extended the block design with black and white patterns, which adds complexity due to the demand for matching orientation. An overview of the task can be seen in figure 4.

To pick or place pieces, users have to point the robot's tip towards and close to the desired location and pull/release a trigger in the handle. The position of the robot and its tip is measured via a motion tracking system[1]. Another button in the handle allows the user to rotate a grabbed piece. Participants are asked to solve the task swiftly and it is completed when all model pieces are copied. Throughout the task execution, we kept track of the user's eye gaze using a robot-mounted remote eye tracker in combination with a 3D gaze model from [3]. Figure 1 shows an example of a participant solving the puzzle.

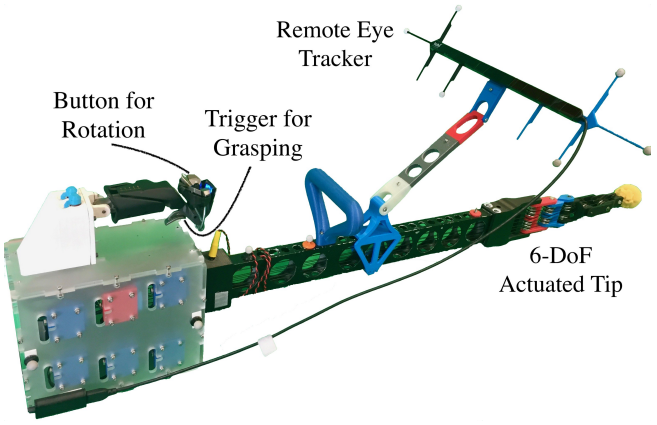



Fig. 3: The handheld robot used in our study. It features a set of input buttons and a trigger at the handle, a 6-DoF tip and user perception through gaze tracking as reported in [3].

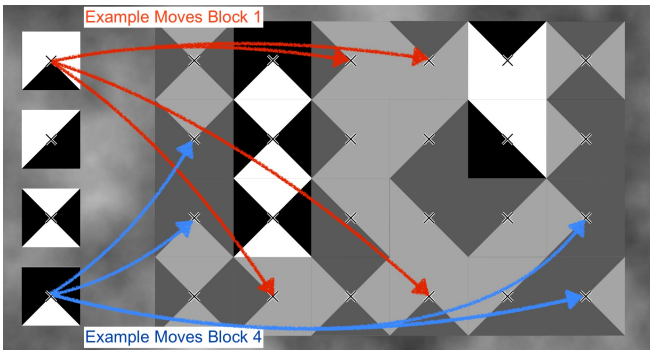



Fig. 4: This shows the layout of the block copy task on a TV display and examples of possible moves for block 1 and 4. Using the robot, a piece from the stock (left column) has to be moved to an associated piece in the pattern (shaded blocks) and match the model's orientation to complete it.

For the data collection, 16 participants (7 females, $m_{age}$ = 25, $SD$ = 4) were recruited. Each completed one practice trial to get familiar with the procedure, followed by another three trials for data collection, where stock pieces and model pieces were randomised before execution. The pattern consists of 24 parts with an even count of the 4 types. The task starts with 5 pre-completed pieces to increase the diversity of solving sequences leaving 19 pieces to be completed by the participant. That way, a total amount of 912 episodes of picking and dropping were recorded.

### B. User Intention Model

In the context of our handheld robot task, we define intention as the user's choice of which object to interact with next i.e. which stock piece to pick and on which pattern field to place it.

Based on our literature review, our model design is guided by the following assumptions:

- **A1** An intended object attracts the users' visual attention prior to interaction.
- **A2** During task planning, the users' visual attention is shared between the intended object and other (e.g. subsequent) task-relevant objects.

Moreover, as noted in [1], a mismatch between the robot's plans and the user's intention inflicts user frustration. Hence, with regards to the model's experimental validation (cf. section VI), we also assume that

- **A3** If the predicted intention is the true intention, a robot that rebels against following the predicted goals induces user frustration.

[1]OptiTrack: optitrack.com

Our method is constrained by the assumption that full task knowledge is available to the system. This includes information about task objects' positions and their relationships such as task-specific matching.

As a first step towards feature construction, the gaze information for an individual object was used to extract a visual attention profile (VAP), which is defined as the continuous probability of an object being gazed. Let $\boldsymbol{x_{gaze}}$ be the 2D point of intersection between the gaze ray and the TV screen surface and $\boldsymbol{x_i}$ the 2D position of the $i$-th object in the screen. Then the gaze position can be compared to each object using the Euclidean distance:

$$d_i(t) = ||\boldsymbol{x_{gaze}} - \boldsymbol{x_i}|| \tag{1}$$

As a decrease of $d$ implies an increased visual attention, the distance profile can be converted to a VAP using the following equation:

$$P_{gazed,i}(t) = \exp(\frac{-d_i(t)^2}{2\sigma^2}) \tag{2}$$

Here, $\sigma$ defines the gaze distance resulting in a significant drop of $P_{gazed}$, which is set to 60 mm based on the pieces' size and tracking tolerance. The intention model uses the VAP of the interval $T_{anticipate} = 4\,\text{s}$ before the point in time of the prediction. Due to the data update frequency of 75 Hz, the profile is discretised into a vector of 300 entries (cf. example in figure 5).

The prediction for picking and placing actions was modelled separately as they require different feature sets. As mentioned above, earlier studies about gaze behaviour during block copying [9] and assembly [16] suggest that the eye gathers information about both what to pick and where to place it prior to initialising manual actions. For this reason, we combined pattern and stock information for picking predictions for each available candidate, resulting in the features selection:

- $F_1$ The VAP of the object itself.
- $F_2$ The VAP of the matching piece in the pattern. If there are several, their VAPs are combined using the element-wise maximum function.

This goes in line with our assumptions **A1**, **A2**. Both features are vectors of real numbers between 0 and 1 with a length of $n = 300$. For the prediction of the dropping location, **A2** is not applicable as the episode finishes with the placing of the part, hence why only $F_1$ (a vector with length $n = 300$) is used for prediction.

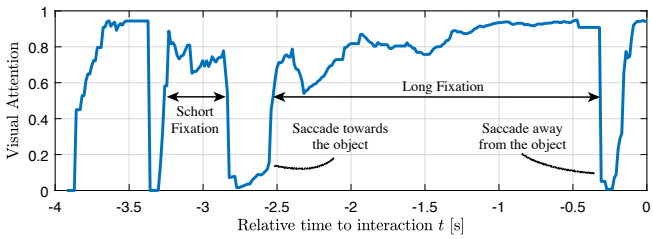



Fig. 5: Illustration of changing visual attention over time within the anticipation window of the prediction model for an individual object.

Note that this feature contains implicit information about fixation durations as well as saccade counts.

An SVM [17] was chosen as a prediction model as this type of supervised machine learning model was used for similar classification problems in the past, e.g. [12]. We divided the sets of VAPs into two categories, one where the associated object was the intended object (labelled as `chosen = 1`) and another one for the objects that were not chosen for interaction (labelled as `chosen = 0`). Training and validation of the models were done through 5-fold cross-validation [18].

The accuracy of predicting the `chosen` label for individual objects is 89.6% for picking actions and 98.3% for placing. However, sometimes the combined decision is conflicting e.g when several stock pieces are predicted to be the intended ones. This is resolved by selecting the one with the highest probability $P(\texttt{chosen} = 1)$ in a one-vs-all setup [19]. This configuration was tested for scenarios with the biggest choice e.g. when all 4 stock parts (random chance = 25%) would be a reasonable choice to pick or when the piece to be placed matches 4 to 6 different pattern pieces (random chance = 17-25%). This validation set $X_{vaild}$ includes 540 picking samples and 294 placing samples. The one-vs-all validation results in a correct prediction rate of 87.9% for picking and 93.25% for placing actions.

## IV. Results of Intention Modelling

The analysis of the intention model's performance is divided into two parts, a quantitative analysis and a qualitative assessment.

### A. Quantitative Analysis

Having trained and validated the intention prediction model for the case where VAPs range over $T_{anticipate}$ prior to $t_0$, the time of interaction with the associated object, we are now interested in knowing to what extent the intention model predicts accurately at some prior time $t_{prior} < t_0$. To answer this question, we extend our model analysis by calculating a $t_{prior}$-dependent prediction accuracy where respective predictions are based on data from the time interval $[t_{prior} - T_{anticipate}, t_{prior}]$. Within a 5-fold cross-validation setup, $t_{prior}$ is gradually decreased while predictions are calculated using the trained SVM model and compared against the ground truth at $t_0$ to determine the accuracy. The validation is based on the aforementioned set $X_{vaild}$ so that the random chance of correct prediction would be $\leq 25\%$. The shift of the anticipation window over the data set is done with a step width of 1 frame (13 ms). This is done for both the case of predicting which piece is picked up next as well as inferring intention concerning where it is going to be placed. For the time offsets $t_{prior} = 0$, 0.5 and 1 seconds, the prediction of picking actions yields an accuracy $a_{pick}$ of 87.94%, 72.36% and 58.07%. The performance of the placing intention model maintains a high accuracy

over a time span of 3 s with an accuracy $a_{place}$ of 93.25%, 80.06% and 63.99% for the times $t_{prior}$ = 0, 1.5 and 3 seconds. In order to interpret these differences in performance, we investigated whether there is a difference between the mean duration of picking and placing actions. We applied a two-sample t-test and found that the picking time (mean = 3.61 s, $SD$ = 1.36 s) is significantly smaller than the placing time (mean = 4.65 s, $SD$ = 1.34 s), with $p < .001, t = -16.12$.

As the prediction model of the picking actions implements the novel aspect of adding the VAPs of related objects, its comparison to existing methods is of particular interest. Figure 6 shows a comparison of the proposed model (where both features $F_1$ and $F_2$ are used) against the case where $F_1$ is the single basis for a prediction, such as the model recently explored in [12]. It can be seen that both models well exceed the chance of picking randomly. Notably, the proposed model outperforms the existing one shortly after the subject ends the preceding move and presumably starts planning the next one. To further investigate the effect of the chosen model on the prediction performance, a two-factorial ANOVA was applied where the prediction time $t$ relative to the action and the model were set as the independent factors and the performance as dependent variable, which reveals that the correct prediction rate of the proposed model is significantly higher ($p < .001$) than the one of the existing model.

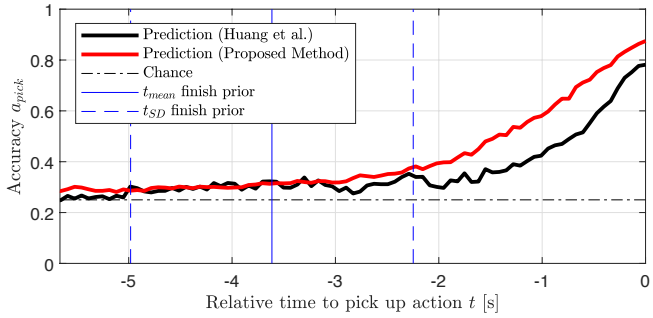



Fig. 6: This diagram shows the performance of predicting pick up actions averaged over 912 samples for two models: our proposed model (red) and an SVM (black), which is based on the feature $F_1$ only, such as proposed by Huang et al. [12]. It can be seen how both models perform better than chance (dashed black) and predict the actions with increasing accuracy as the prediction time $t$ approaches the time of the action's execution $t = 0$. $t_{mean}$ (with temporal SD $t_{SD}$) is the mean time of completing the last block and hence the earliest meaningful time of predicting picking as a subsequent action.

### B. Qualitative Analysis

For an in-depth understanding of how the intention models respond to different gaze patterns, we investigate the prediction profile i.e. the change of the prediction over time, for a set of typical scenarios.

*1) One Dominant Type:* A common observation was that the target object perceived most of the user's visual attention before interactions, which goes in line with our assumption **A1**. An example of these *one type dominant* samples can be seen in figure 7a. A subset of this category is the case where the user's eye gaze alters between the piece to pick and the matching place in the pattern i.e. where to put it (cf. figure 7b), which supports our assumption **A2**. For the majority of these one type dominant samples both the picking and placing prediction models predict correctly.

*2) Trending Choice:* While the anticipation time of the pick-up prediction model lies within a second and is thus rather reactive, the placing intention model is characterised by a slow increase of likelihood during the task i.e. it shows a low-pass characteristic. Figure 8 demonstrates that the model is robust against small attention gaps and intermediate glances at competitors, however, the model requires an increased time window to build up confidence.

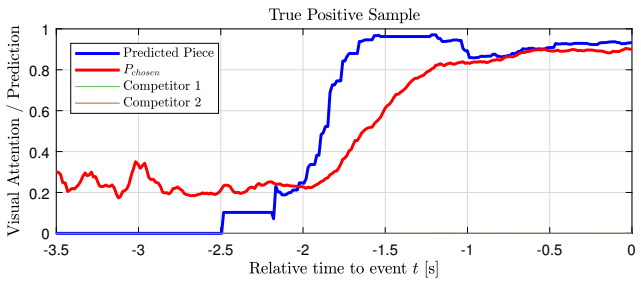
*3) Incorrect Predictions:* There is a number of reasons for an incorrect prediction. Most commonly, a close-by neighbour received more visual attention and was falsely classified as the intended object. In other cases, it was impossible to predict the intended object using our model due to missing saccades towards it or faulty gaze tracking.

## V. Discussion of Intention Modelling

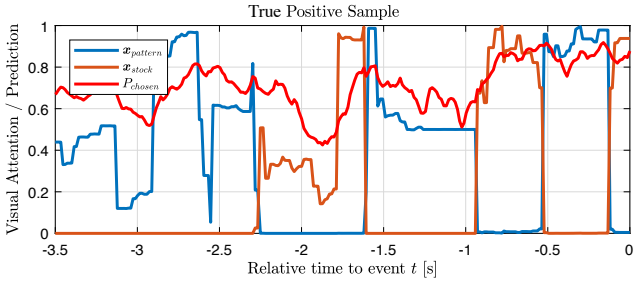
In addressing research question **Q1**, we proposed a user intention model based on gaze cues for the prediction of actions, which was assessed in a pick and place task. As a novel aspect introduced through this study, the predictions are not only based on saccades and fixation durations of an individual object but also on those of related objects. In other words, assessing the attention on objects in the workspace helps to predict which piece outside the current workspace is needed next. When the subject turns his/her attention towards the piece, the model interprets this as a confirmation rather than the start of a selection process. This helps to cut the time required for the model to gather relevant gaze information and makes predictions more reliable than traditional models.

We showed that, within this task, the prediction of different actions has different anticipation times i.e. the model allows predictions 500 ms before picking actions (71.6% accuracy) and 1500 ms prior to dropping actions (80.06% accuracy). This can partially be explained by the fact that picking episodes are shorter than placing episodes. More importantly, we observed that users planned the entire pick-place cycle rather than planning picking and placing actions separately. This becomes evident through the qualitative analysis, which shows altering fixations between the piece to pick and where to place it. That way, the placing prediction model can already gather information at the time of picking.

In terms of the system's limitations, we point out that it is unclear how well the model generalises and per-

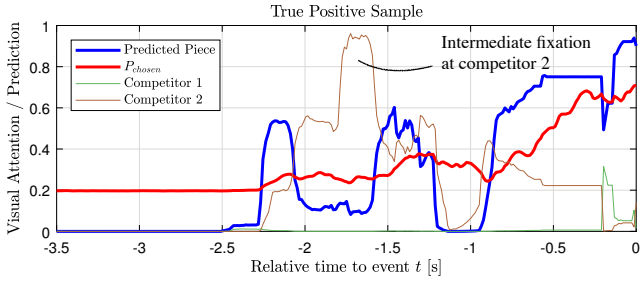



(a) One piece receives most of the user's visual attention prior to placing




(b) User gaze alters between stock piece and matching workspace location

Fig. 7: These diagrams show examples of correct predictions for *one type dominant* samples. (a) shows, how long fixation times (blue) results into a high probability value (red) e.g. for a location to place a piece. Similarly, (b) shows, how the prediction model links the VIPs of related objects. The subject's gaze alters between two related objects e.g. a piece to pick up and a matching location to place it (cf. orange and blue VAPs) leading to a high probability estimation (red) for this piece being the user-intended one.



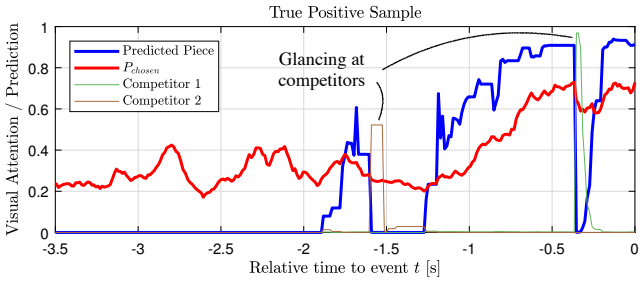




Fig. 8: These two examples illustrate how the visual attention (blue) of an object builds up during the user's decision process in which case the intention prediction (red) remains undecided ($P_{chosen} < 0.5$) for a longer time compared to the case where no competition receives fixations (cf. fig 7).

forms for new tasks as the differences in performances for the two example actions (picking and dropping) indicate that the model is task-dependent.

The results are encouraging for testing the prediction model in a real-time application. Therefore, we proceed with an experimental study where the intention model is used for cooperative behaviour.

## VI. Intention Prediction Model Validation

In the second part of our study, we validate the proposed intention model for the case where it is used to control the robot's behaviour and motion. While the aforementioned experiments and analysis demonstrate that the intention model is capable of predicting users' short term goals while having full control over the robot's tip, it is unclear whether this is true for the case where the robot reacts to these predictions. For example, users might adapt their intention to the robot's plans just by seeing it moving towards a target that might differ from their initially intended move. That way, labelling the robot's predictions as being correct or incorrect in the same way as we did in the first study becomes invalid due to the lack of ground truth. For this reason, we propose to assess the intention model indirectly instead by observing users' reactions to the predictions with a focus on frustration. Therefore, we base our experimental validation on assumption **A3** and use frustration as a measure of correct predictions.

### A. Intention Affected Robot Behaviour

For the experimental validation of the intention model, we used the aforementioned block copy task and introduced an assistive behaviour to the robot, which is controlled based on the predictions of a user's intended subsequent move. We created 3 different behaviour modes: *Follow Intention*, *Rebel* and *Random*. Where we note that rebellion itself, as a mode of operation, has been argued as a useful concept for constructive purposes [20]. For each mode, the robot retreats to a crouched position while there is a low probability for each available target. When the probability of the target with the highest probability reaches a threshold, the robot reacts as follows:

- **Follow Intention:** The robot moves towards the target with the highest predicted intention.
- **Rebel:** The robot avoids the target with the highest prediction and moves towards the target with the lowest predicted intention instead.
- **Random:** The robot choses a random valid target.

As per assumption **A3**, we argue, that an observed reduction of user frustration in the *Follow Intention* mode compared to the *Rebel* mode would validate that the predicted user intention went in line with the true intention. A demo of the behaviour modes can be seen in the supplementary video of this paper and on our webpage [8].

### B. Experiment Execution

We recruited 20 new participants (6 females, $m_{age}$ = 26, $SD$ = 4) for the validation study of which 2 were later removed from the set for data analysis due to malfunctioning gaze tracking. Each was asked to first complete the task without the robot moving for familiarisation with the rules and the robot handling. This practice session was followed by 3 trials where, for each, the robot's behaviour was set to a different behaviour mode. The block pattern to complete as well as the order of the behaviour modes were randomised. Furthermore, 5 (out of 24) randomly chosen blocks were pre-completed to stimulate some diversity in solving strategies e.g. to prevent repeated line-by-line completion.

The participants were told to solve the trial tasks swiftly and that their performance was recorded. They did not receive any information about the behaviour modes, but were told that the robot will move and try to help them with the task. Each trial was followed by the completion of a NASA Task Load Index (TLX) form [21] and 3 min resting time.

## VII. Results and Discussion: Model Validation

To determine the effect of the robot's behaviour mode on the subjects' frustration level, we performed an analysis of variance (ANOVA) with the mode as the independent variable and the frustration component of the TLX as a dependent variable. As the analysis yielded a significant effect ($p = .023$), it was further explored using a post-hoc pairwise t-test with applied Bonferroni correction. The frustration mean for the *Rebel* group was identified as being significantly higher than in the *Follow Intention* group ($p = .019$). No significant mean differences were found when comparing the *Random* group to the others. The results can be seen in table 1 and figure 9.

We extended our analysis to both, the combined TLX results, which serve as an indicator for perceived task load, and the measured performance, which is defined as the number of completed blocks per minute. However, an applied ANOVA did not yield an effect of the robot's behaviour mode, neither on the combined TLX nor on the performance.

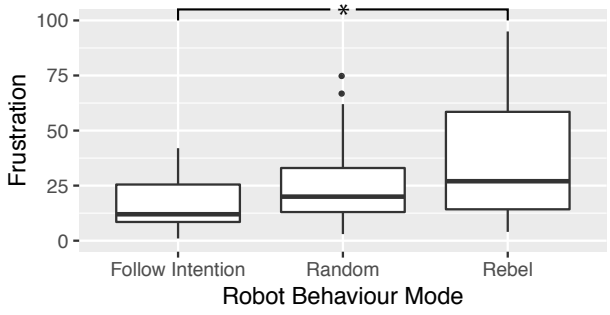



Fig. 9: Perceived frustration from the TLX results for each of the tested behaviour modes. The mean values of starred groups yield a significant difference (cf. table 1).

| | **Follow Intention** | **Random** |
|---|---|---|
| **Rebel** | $p = .019$ * | $p = .495$ |
| **Random** | $p = .469$ | - |

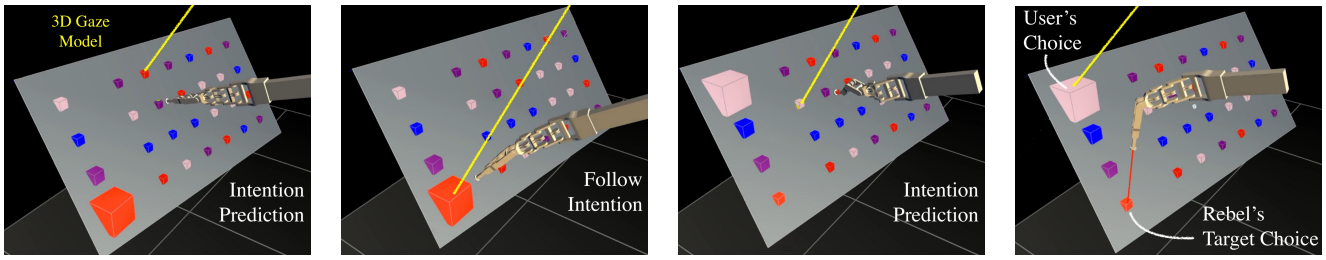
TABLE 1: Bonferroni corrected $p$-values of pairwise t-test results for the differences in mode depended frustration means. The starred value is significant ($p < .05$).

As part of a qualitative review of the robot's behaviour we found that in the *Rebel* mode, participants perform an increased number of corrective moves compared to the *Follow Intention* scenario. Figure 10 shows how the robot's aim matches the user's intention in the *Follow Intention* mode whereas in the *Rabel* example, the user rushes towards the intended aim but needs to correct his move as the robot aims for a different piece.

Some participants commented on the behaviour modes. The *Follow Intention* mode was often preferred (e.g. "I liked being in charge and the robot was helpful" and "The robot followed my decisions") whereas the *Random* mode lead to irritation in some users (e.g. "First I thought it would go where I wanted but then it started moving unpredictably"). For the *Rebel* mode, we observed divergent reactions. While some subjects struggled because of the mismatch between the robot's motion and their plans, others started following the robot's lead. This was also reflected in the comments e.g. "Now the robot does its own thing, I don't like it" versus "It was easier because I did not have to think much".

The observed difference in frustration ratings between the mode where the robot supports the user's predicted intention versus avoiding it is evidence for most of the intention predictions matching the true intention. This validates the proposed intention model and its application in assisted reaching. With regards to **Q2**, our interpretation of the results is that during the *Follow Intention* trials, the robot did follow the users' preferred sequence rather than the users adapting it to the robotic motion. That way, the intention model enhances cooperation concerning action anticipation between collaborators. While this is an important cooperation characteristic, there are more layers to it such as intention communication and the adaptation to other user preferences, which leaves space for future exploration.

Mean frustration for the *Random* mode being between the other two modes is expected, given the robot's choices contain both predicted and non-predicted targets. The effect size is too small for a reliable distinction within this group size. Our analysis furthermore shows that user frustration is more sensitive to the robot's intention prediction than perceived task load and performance. Therefore, we suggest that collaborative agents should follow user intention when there are subtasks with similar priorities for enhanced cooperation.




(a) Prediction of the red piece during placing of the purple piece.

(b) The robot's motion goes in line with the user's intention as it adapts its plans.

(c) Prediction of the pink piece while placing the purple one.

(d) Avoiding user intent leads to a mismatch with the user's tactical motion.

Fig. 10: These figures illustrate the systems' underlying intention estimation and how the different modes affect cooperation. The users' eye gaze model is represented as a yellow line while the estimated probability for a piece to be chosen by the user is indicated by its size. It can be seen how following the intention prediction assists the user with his/her choice (a,b) while avoiding the intended object (c,d) forces the user to adapt his/her plan to the robot's motion.

## VIII. Conclusion

We investigated the use of gaze information to infer user intention within the context of a handheld robot. A pick and place task was used to collect gaze data as a basis for an SVM-based prediction model. Results show that depending on the anticipation time, picking actions can be predicted with up to 87.94% accuracy, 500 ms ahead and dropping actions with an accuracy of 93.25%, 1500 ms ahead. We show that merging gaze information with respect to objects that are linked to the same task in a single model helps to increase the prediction performance. The introduction of frustration via rebellion was used to measure the usually complex aspect of effectiveness of intention prediction in human-robot interaction. An approach that, together with the model proposed, could be useful in other cooperative robot studies.

### *Acknowledgements and Data Access*

Data from this study is available by contacting the authors. Thanks to the German Academic Scholarship Foundation and UK's EPSRC for funding. Stated opinions are the authors' and not of the funders.

## References

[1] Austin Gregg-Smith and Walterio W Mayol-Cuevas. The design and evaluation of a cooperative handheld robot. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1968–1975. IEEE, 2015.

[2] Austin Gregg-Smith and Walterio W Mayol-Cuevas. Investigating spatial guidance for a cooperative handheld robot. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3367–3374. IEEE, 2016.

[3] Janis Stolzenwald and Walterio Mayol-Cuevas. I Can See Your Aim: Estimating User Attention From Gaze For Handheld Robot Collaboration. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3897–3904, July 2018.

[4] Chien-Ming Huang and Bilge Mutlu. Anticipatory robot control for efficient human-robot collaboration. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 83–90. IEEE, 2016.

[5] Harish chaandar Ravichandar and Ashwin Dani. Human intention inference through interacting multiple model filtering. In *2015 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, 2015.

[6] Tingting Liu, Jiaole Wang, and Max Q H Meng. Evolving hidden Markov model based human intention learning and inference. In *2015 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 206–211. IEEE, 2015.

[7] Austin Gregg-Smith and Walterio W Mayol-Cuevas. Inverse Kinematics and Design of a Novel 6-DoF Handheld Robot Arm. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2102–2109. IEEE, 2016.

[8] Handheld Robotics: handheldrobotics.org.

[9] Dana H Ballard, Mary M Hayhoe, and Jeff B Pelz. Memory Representations in Natural Tasks. *Journal of Cognitive Neuroscience*, 7(1):66–80, 1995.

[10] Harish chaandar Ravichandar, Avnish Kumar, and Ashwin Dani. Bayesian Human Intention Inference Through Multiple Model Filtering with Gaze-based Priors. In *th International Conference on Information Fusion FUSION*, pages 2296–2302, June 2016.

[11] Hema S Koppula and Ashutosh Saxena. Anticipating Human Activities Using Object Affordances for Reactive Robotic Response. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):14–29, 2016.

[12] Chien-Ming Huang, Sean Andrist, Allison Sauppé, and Bilge Mutlu. Using gaze patterns to predict task intent in collaboration. *Frontiers in Psychology*, 6(1049):1–12, July 2015.

[13] Michael Land, Neil Mennie, and Jennifer Rusted. The Roles of Vision and Eye Movements in the Control of Activities of Daily Living. *Perception*, 28(11):1311–1328, 1999.

[14] Michael F Land and Mary Hayhoe. In what ways do eye movements contribute to everyday activities? *Vision Research*, 41(25-26):3559–3565, November 2001.

[15] Roland S Johansson, Göran Westling, Anders Bäckström, and J Randall Flanagan. Eye–Hand Coordination in Object Manipulation. *Journal of Neuroscience*, 21(17):6917–6932, September 2001.

[16] Neil Mennie, Mary Hayhoe, and Brian Sullivan. Look-ahead fixations: anticipatory eye movements in natural tasks. *Experimental Brain Research*, 179(3):427–442, December 2006.

[17] M A Hearst, S T Dumais, E Osuna, J Platt, and B Scholkopf. Support Vector Machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28, 1998.

[18] Ron Kohavi. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *International Joint Conference on Artificial Intelligence IJCAI*, pages 1137–1145, 1995.

[19] Ryan Rifkin and Aldebaro Klautau. In Defense of One-Vs-All Classification. *Journal of Machine Learning Research*, pages 101–141, June 2004.

[20] David W Aha and Alexandra Coman. The AI rebellion: Changing the narrative. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, 2017.

[21] Sandra G Hart and Lowell E Staveland. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, pages 139–183. Elsevier, 1988.