Communicative Cues for Reach-to-Grasp Motions: From Humans to Robots

Robotics Track

Doğancan Kebüde, Cem Eteke, Tevfik Metin Sezgin, Barış Akgün
Koç University
Istanbul, Turkey
{dkebude16,ceteke13,mtsezgin,baakgun}@ku.edu.tr

ABSTRACT

Intent communication is an important challenge in the context of human-robot interaction. The aim of this work is to identify subtle non-verbal cues that make communication among humans fluent and use them to generate intent expressive robot motion. A human-human reach-to-grasp experiment ($n = 14$) identified two temporal and two spatial cues: (1) relative time to reach maximum hand aperture (*MA*), (2) overall motion duration (*OT*), (3) exaggeration in motion (*Exg*), and (4) change in grasp modality (*GM*). Results showed there was statistically significant difference in the temporal cues between no-intention and intention conditions. In a follow-up experiment ($n = 30$), reach-to-grasp motions of a simulated robot containing different cue combinations were shown to the participants. They were asked to guess the target object during robot's motion, based on the assumption that intent expressive motion would result in earlier and more accurate guesses. Results showed that, *OT*, *GM* and several cue combinations led to faster and more accurate guesses which imply they can be used to generate communicative motion. However, *MA* had no effect, and surprisingly *Exg* had a negative effect on expressiveness.

KEYWORDS

human-robot interaction; motion legibility; communicative cues

ACM Reference Format:
Doğancan Kebüde, Cem Eteke, Tevfik Metin Sezgin, Barış Akgün. 2018. Communicative Cues for Reach-to-Grasp Motions: From Humans to Robots. In *Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), Stockholm, Sweden, July 10–15, 2018,* IFAAMAS, 9 pages.

1 INTRODUCTION

Recent advances in robotics points towards a new horizon: Cage-free robots working alongside humans in shared workspaces. To achieve fluency and efficiency in such scenarios, partners need to understand each other's intent. Humans are very good at intent expression in the context of joint action [20]. However, expressing intent is an ongoing challenge for human-robot interaction (HRI). For the purposes of this paper, we are interested in intent to act on an object vs. another in reach-to-grasp motions, as in Fig. 1.

Several studies tried to solve the intent expression problem for robots; however, very few of them looked at what makes non-verbal communication among humans fluent. In the human-human cases, fluency can be achieved naturally by verbal and non-verbal subtle cues that clarify the motion intent. We are interested in the subtle non-verbal cues of human motion that are not recognizable without careful consideration. These cues can be temporal, spatial, morphological or force related.

Figure 1: A participant reaching to grasp one of the cylindrical objects.

We argue that these subtle cues can be utilized by robots to express intent which will increase fluency of human-robot teams. In order to identify these subtle cues, we explore human-human interaction (HHI) scenarios in shared workspaces and collaborative activities. There is a gap between the robotics studies and HHI studies in the context of intent expression. We think that humans are a valuable source of information for intent communication.

For expressing robot's intent, multiple studies have looked at communicative action generation. In [9], authors defined the differences between "legible" and "predictable" motions and showed that legible motion can improve human recognition of robot's intent. They investigated optimal legible motion generation in [10]. They also showed how legible motion improves collaboration and how functional motion may disrupt collaboration fluency in [8]. The work in [22] defined "simple", "curved" (defined as"predictable" and"legible" respectively by [9]) and "straight" motions. The study showed that straight motions improve intent expression more than their curved counterparts. Finally, the motions defined by [9] and [22] are compared in [6] to find that there is no optimal solution to the intent expression problem for fluent collaboration yet. [6] also showed that simple and straight motions can be as expressive as the legible motion depending on robot's anthropomorphism. These studies do not directly use human data. Dragan et al. [9] takes inspiration from cognitive studies to formulate the notions of legibility and predictability. However, how much exaggeration to produce is not clear and whether humans actually use exaggeration

Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), M. Dastani, G. Sukthankar, E. André, S. Koenig (eds.), July 10–15, 2018, Stockholm, Sweden.

to express intent is an open question. In this paper, we refer to any motion generated in a shared workspace scenario as *communicative*.

All of the aforementioned studies investigated and implemented spatial aspects of motion. We think that temporal aspects of motion are important as well for communicating intent. Gielniak and Thomaz [13] looked at spatio-temporal correspondence (STC) in human motion. Using STC as a metric to optimize the temporal aspects of motion in addition to spatial aspects improved (1) recognition of motion as a common human motion, (2) accuracy of intent identification and (3) accuracy of mimicking. Although not directly utilizing HHI studies, this study shows that looking at human motion can be helpful in designing intent expression.

There are several relevant studies that investigate communicative motion cues in collaborative human-human interactions in the joint action psychology literature such as: identifying the effect of social intention on motion [4], stating that the information encoding motion intent is available in motion kinematics [3], and demonstrating that this information can be inferred from some cues in the movement [19]. All of these studies investigated reach-to-grasp motions and designed their experiments accordingly.

The effects of conscious, but not exaggerated, effort to make the action communicative is investigated in [18]. The idea is that the subtle cues would be amplified by such an effort. They demonstrated, in two separate experiments, that there are subtle cues in reach-to-grasp motions that make the action communicative such as deviation from the non-communicative trajectory, time of maximum hand aperture and time of peak closing velocity.

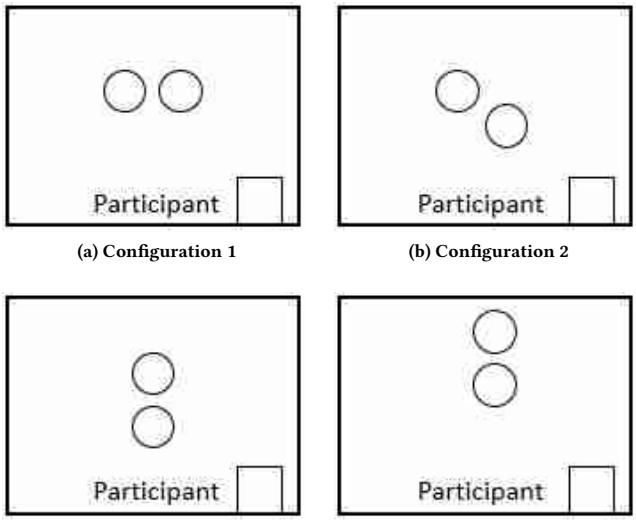
This prompted us to do a similar experiment to the one in Sartori et al. [18] from an HRI perspective. Our aim was to replicate some of their results and to identify additional cues that robots can utilize. We modified Sartori et al. [18]'s experiment to be conducted in a less intrusive environment for the participant by removing marker based motion tracking. We identified additional cues to measure (see Sect. 2.3.2). We hypothesized that these cues would improve fluency and communication in HRI if used in robot motion generation. We tested this by generating robot motion using these cues in a simulated environment with a second experiment. Participants were motivated to guess the object that the robot is reaching to as early and as accurately as possible. We show that most cues and cue combinations have an effect on intent expressiveness.

2 HUMAN-HUMAN EXPERIMENT

In the first experiment, our aim is to identify the intent expressive cues in human-human interaction. We investigate how humans reach to grasp a cylindrical object from a table, as was done in Dragan et al. [9] by a robot. This is done with another cylindrical object in close vicinity, as seen in Fig. 1. The experiment had two conditions; with and without an observer present. The study followed a within-subject design, i.e., all the participants performed the experiments in both conditions in a counterbalanced order.

2.1 Experimental Setup

The experimenter and the participant take their places on opposite sides of a table in a room where a video camera is set to record the motions. The camera is placed diagonally, to avoid influencing the participant, depending on participant's handedness (i.e. if participant is right-handed, then the camera is placed at his/her right



(a) Configuration 1 (b) Configuration 2

(c) Configuration 3 (d) Configuration 4

Figure 2: Top-view for the configurations of the objects when the participant is right handed and sits facing the bottom edge of each figure. Square at the bottom right shows the initial hand position.

diagonal and vice versa). Two cylindrical bottles are placed on the table, 5 cm apart from each other, within participant's reach. One of the bottles is filled with water, and the other with tea (see Fig. 1).

The participants are asked to reach, grasp, take and put back one of the objects in four configurations. These four configurations are: (1) objects reside side-by-side (Fig. 2a), (2) one of the objects resides diagonally in front of the other, partially blocking participant's reaching motion (Fig. 2b), (3) one of the objects is directly in front of the other, partially blocking participant's view of the other object (Fig. 2c), (4) one of the objects is directly in front of the other, at the edge of participant's reach (Fig. 2d). These are designed to partially block the straight line motion and/or vision to affect participant's action. The fourth configuration was designed to check if there are any changes in participant's motion between the cases where the object is easy-to-reach vs. when it requires some extra effort.

The experiment starts with informing the participants. Then, they are asked whether they are left or right handed to setup the environment. They are also instructed to start from a predetermined initial position and to always reach with the same hand. The experiment proceeds based on the counterbalance order. In each object configuration, the participant repeats the action twice per object and for each configuration, the participants follow the order of objects to reach to, based on a list provided on a piece of paper.

2.1.1 Control Case. The participants reach the bottles when the experimenter is sitting at the opposite side of the table. The experimenter pretends to be working on his laptop. He does not look at the participant or engage with him/her in any way. The participants signal the experimenter when they are done with a specific configuration. The experimenter then switches to the next configuration and goes back to working on his laptop.

2.1.2 Communicative (Test) Case. In this case, the experimenter acts as an observer. The participants reach the bottles when the

observer is sitting at the opposite side of the table but this time paying attention to the participants' actions. In addition, *the participants are explicitly asked to communicate their intent* (i.e. which object they are reaching to) as understandable as possible by only the motion itself (no gaze, no pointing, no verbal communication). Observer does not provide any feedback to the participant, the interaction is completely non-verbal.

For counterbalancing, half of the participants start the experiment with the control case while others start with the test case.

2.2 Pilot Phase

5 participants (4 male and 1 female) attended the pilot phase of the experiment. Two of them were already familiar with the work and thus their communicative motions were very similar to [9]'s legible motion definition. However, the pilot phase was still useful for identifying potential subtle communicative cues of reach-to-grasp motions. Furthermore, pilot data was used in an a priori power analysis to decide on the number of participants for the main experiment.

2.2.1 Hypotheses. Following the pilot phase of the experiment, the following hypotheses were established:

(1) Overall motion duration increases in the test case.
(2) Maximum hand aperture occurs earlier in the test case.
(3) Participants might exaggerate their motion in the test case.
(4) Grasp modality might change between configurations, and between control and communicative (test) cases.

2.2.2 A Priori Power Analysis. An a priori power analysis using the relative time to maximum hand aperture (see Sect. 2.3.2) was conducted to decide how many participants would be needed for the experiment. This novel cue was chosen since it was not investigated by Sartori et al. [18]. G*Power 3.1 software [12] was used for power analysis with a calculated effect size of 0.92 based on the data collected in the pilot phase, α error probability of 0.05 and for a power of 0.90. The power analysis resulted in an expected power of 0.91 for a sample size of 12. To be on the safe side, the experiment was performed with 14 participants.

2.3 Experiment Phase

2.3.1 Participants. 7 male and 7 female participants attended the experiment (median age 21.5, all from a university campus community). 9 of them stated that they had no robotics experience at all, 4 of them stated that they watched some robot videos online or had minor interactions with robots and only one of them stated that he workd with robots but not in a research related way.

At the end of the experiment, the participants were asked to complete a survey with the following questions:

(1) How much did you change your motion during the communicative case? (Likert scale question with 5 degrees)
(2) If you changed your motion, what were the key aspects you put importance to?

2.3.2 Data Collection and Analysis. A total of 224 motions were recorded with a camera, from fourteen participants in four configurations reaching each bottle twice. A frame-by-frame analysis of the resulting videos was done to identify the frames at which:

- the motion started: t_{start},
- the hand started opening: t_{open},
- the hand reached its maximum aperture: $t_{aperture}$,
- the hand reached the object (~40% of opposing digits passed the centroid of the bottle): t_{reach},
- the hand started closing: t_{close},
- the hand grasped the object: t_{grasp}.

This process is not as accurate as a marker based one but the time frames are accurate to ±33ms (30 fps), and the setup is less intrusive.

The gathered time stamps were used to calculate three variables:

Overall motion duration t_{total}: It was seen during the pilot phase that the overall motion duration increases during the communicative case.

$$t_{total} = t_{reach} - t_{start} \tag{1}$$

Time to maximum hand aperture t_{max}:

$$t_{max} = t_{aperture} - t_{start} \tag{2}$$

Relative time to maximum hand aperture r_{max}: The time of maximum hand aperture by itself does not seem to be a good variable since the motion speed and overall motion duration change between control and communicative cases. We argue that a variable relative to the overall duration would be more robust in describing maximum hand aperture timing. It was seen during the pilot phase that the maximum hand aperture is reached earlier in the communicative case.

$$r_{max} = \frac{t_{max}}{t_{total}} \tag{3}$$

Other than these numerical variables, there were two qualitative variables that were observed during the pilot study that might affect the the communicative aspects of motion:

Exaggeration The "legible" motion defined by Dragan et al. [9] corresponds to exaggerated gestures. During the experiment, the reach-to-grasp motion for each participant was investigated to identify exaggerated motions such as qualitatively noticeable (~15cm) deviations from the regular trajectory. It was seen during the pilot phase that the participants that had intimate knowledge of the study exaggerated their motions significantly while the others did so slightly and only in the edge cases.

Sartori et al. [18] also stated that deviation from the straight trajectory might be a good indicator for intent-expressive motion. However this deviation was at most 1.6 cm in their experiments which is not measurable from our video recordings. Following that, this study does not measure the actual deviation but only observes if *there is* noticeable exaggeration or not.

Grasp Modality Another observation of the pilot phase was that some participants tend to change their grasp modality between cases (e.g. grasping the bottle from the side during the control case vs grasping the bottle from the top for the communicative case). Based on this, we decided to consider grasp modality as another potential cue for communicative motion.

The previous studies in communicative collaboration worked on spatial components of motion and they did not look at temporal components. The first three variables describe temporal aspects of human motion that can easily be implemented in robotics, for example as constraints to motion planning or explicit targets to spline based trajectory generation. As for the spatial components,

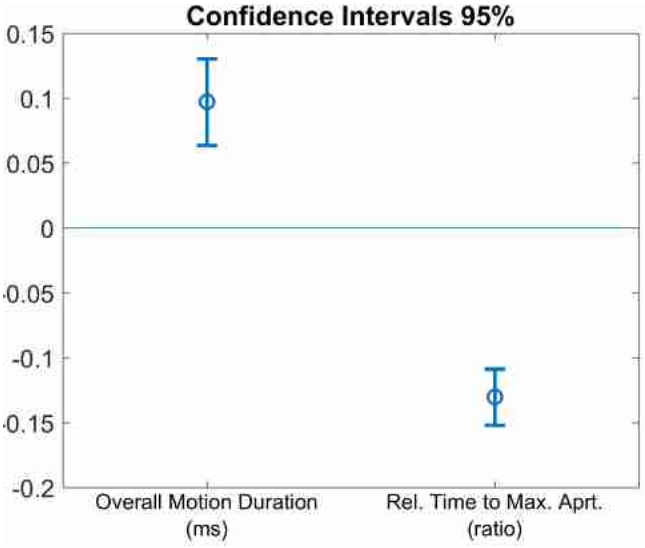


Figure 3: 95% confidence intervals for differences between overall motion durations (ms) and differences between relative time to maximum hand aperture ratios.

exaggeration is already implemented by several studies in robotics. Picking different grasp modalities is also relatively easy to implement. There are potentially other subtle cues that humans exhibit during communicative motion but we chose to investigate the aforementioned five.

2.3.3 Experimental Results. The data gathered from the experiments were analyzed according to the hypotheses described in Sect. 2.2.1 based on the metrics described in Sect. 2.3.2. The participants' responses to the post-experiment survey questions described in Sect. 2.3.1 are also considered in this section.

Overall motion duration t_{total}: Overall motion duration for the communicative case is significantly different than the control case with mean 650 ms in control case vs 750 ms in the communicative case with $p < 1e-5$ as calculated by a cumulative left-tailed (one-way hypothesis) paired t-test. This result supports the first hypothesis in Sect. 2.2.1 that the overall motion duration increases for the communicative case. Fig. 3 provides a 95% confidence interval that shows this increase.

Time to maximum hand-aperture t_{max}: This variable is calculated as it was also used in [18]. Since the motion duration tends to change between control and communicative cases, we argue that the actual time to maximum hand aperture can be misleading.

Sartori et al. [18]'s findings showed that the maximum hand aperture is achieved later in the communicative case. However, our study showed that this might change with the motion duration. Our results did not show any significant difference between the control case and the test case for maximum hand-aperture timing (mean 463 ms. for the control case vs. 430 ms. for the communicative case). This discrepancy between our study and [18] might be caused by cultural differences between the participants or the differences between the experimental task and setup.

Relative time to maximum hand aperture r_{max}: The relative time to maximum hand aperture gives us a more robust clue as to whether hand aperture timing is a communicative cue or not. The experimental results show that there is a significant difference between the control case and the communicative case in terms of this variable (mean 71% in the control case vs. 59% in the communicative case). This means *the maximum hand aperture occurs much earlier in motion when a conscious communicative effort is present.*

For all iterations and all subjects, a cumulative left-tailed (one-way hypothesis) paired t-test results in a p-value well below 1e-5, supporting the second hypothesis in Sect. 2.2.1. Fig. 3 provides a 95% confidence interval that shows the relative time to maximum hand aperture decreases in the communicative case.

Exaggeration: As opposed to the trajectories generated by [9], we did not observe much exaggeration in communicative motions of humans. There might be less than 1.6 cm of deviation as Sartori et al. [18] measured, which is not observable from the video recordings, contradicting the third hypothesis presented in Sect. 2.2.1. Still, exaggeration occurred for some participants in some edge cases. This happened in configurations 2, 3 and 4, in which one of the bottles partially block the other one visually/motion-wise (Fig. 2). This does not imply that exaggeration is not a useful communicative cue for robot action generation as others in the field of robotics showed that it is useful in communicative motion. In addition, exaggeration is used extensively in animation [15].

Grasp Modality: Grasp modality proved to be an important cue for communicative action for certain participants. While the participants did not care much about how they grasped the objects in the control case, they tend to change their grasp modality in the communicative case. For sixteen reach-to-grasp motion comparisons between control and test cases and for all the participants, 103 grasp modality changes were observed out of 224 motions. Most of the participants decided that the object that is closest to their hand should be grasped from the front or the side while the object that is further to them should be grasped from either the top or the opposite side.

The change of grasp modality also caused some sort of exaggeration. This was not in terms of the whole motion trajectory but in terms of how the wrist bends while the object is being grasped.

Participant Responses to Survey Questions: For the first question, only one participant responded as "changed significantly", where 6 of the participants responded as "changed a little", 6 others responded as "did not change much" and one last participant responded as "did not change at all".

An important observation from the responses is that, the participants were not aware of the relative time to maximum hand aperture, which significantly changed for all of them. None of them reported a conscious change in their timing, i.e, *hand aperture timing, might be one of the key non-verbal communicative cues in HHI.*

For the second question, most of the participants provided responses that support the subtle cues observed in this study. Some of the participants responded that they decreased their motion speed, while some others said that they changed their grasp modality.

Another interesting aspect which is common in four of the participants' responses is that they changed their body posture and oriented their chest towards the object they were reaching to. It can be inferred from these responses that, observations for communicative motion should not only check the motion kinematics of the arm but also the body orientation and posture.

2.3.4 Post Hoc Power Analysis. A post hoc power analysis using the relative time to maximum hand aperture was conducted to see if the experiments could achieve the desired target power of 0.9. G*Power 3.1 software [12] was used to conduct the post hoc power analysis with an effect size of 0.78 calculated through G*Power with the data collected in the experiment phase, α error probability of 0.05 and for a sample size of 14 and it resulted with an actual power of 0.87. Since this is very close to the a priori power analysis and the calculated power, we decided not to include further participants.

3 SIMULATED ROBOT EXPERIMENT

The goal of the first experiment was to identify subtle commnicative motion cues in humans that can be reproduced by robots. We conlude that this goal was reached based on the positive results. The next step is to test whether these cues are actually useful in robotic communicative motion generation. Before going further with an experiment that involves a real robot, the second experiment was conducted with videos recorded via simulation to be able to systematically test a large set of cue combinations. The following four cues from the previous experiment were utilized: relative time to maximum hand aperture (*MA*), overall motion duration (*OT*), exaggeration (*Exg*) and grasp modality (*GM*).

The second experiment had a $2 \times 2 \times 2 \times 2$ multi-factorial design based on the cues being present or not. The case when no cue is present corresponds to the no intention case (*Non*). The resulting cue combinations were used to generate robot motion for three object configurations of the first experiment (see Sect. 2.1). The fourth one was not investigated since no specific change between the third and fourth configurations was observed in the first experiment.

Robot trajectories were generated with a spline-based approach: The timings and waypoints, which result in the cues, were chosen according to one of the first experiment participants' motions that has shown all of the identified communicative cues. Then, the chosen waypoints were utilized to decide on parameter values of a third order spline based trajectory for the end effector. In each configuration, the robot reached both of the objects. This resulted in $2^4 \times 3 \times 2 = 96$ robot reaching-to-grasp trajectories.

The participants were shown the trajectories and asked to identify the object the robot was reaching to grasp as soon as they were sure. They were encouraged to maximize a score that is based on both accuracy and speed, as described in Sect. 3.3 with Eq. 7.

Screenshots from two example trajectories for the same configuration and the same object can be seen in Fig. 4. The top row depicts a trajectory with no cues (*Non* case) and the bottom depicts a trajectory with *MA*, *OT* and *Exg*. The latter takes longer and reaches its maximum aperture earlier relative to the overall duration.

3.1 Experimental Setup

The experiment is conducted via a graphical user interface (GUI). The main function of the GUI is to display the simulated robot videos to the participants and record their object guess. Screenshots from the GUI can be seen in Fig. 4 and Fig. 5. All the participants use the same computer that runs the GUI. The GUI starts with a form to collect participant's name, id, gender, age and robotic experience. Then, the experiment is carried out in three steps; (1) Preview, (2) Tutorial and (3) Actual Experiment. During the Tutorial and the Actual Experiment, participants hit the space bar to play the next video. They hit it again to stop the video when they are sure about the robot's target. Then, they are shown two buttons to make their guess, as depicted in Fig. 5a. The specifics of the steps are as follows:

Preview: The participants are shown six videos of the simulated robot while it is not expressing intent (*Non*). This step is for the participants to get familiar with the robot's movement so that they will only be concerned with robot's intent in the later steps . The participants are not asked to understand the robot's intent.

Tutorial: The GUI shows six videos from the *Non* case and six videos from the *MA + OT + Exg* case, i.e. when these three cues are active, at random. This step is for the participants to get familiar with the GUI and the experiment. To motivate the participants in making better guesses during the actual experiment, the calculated participant score is shown at the end of this step.

Actual Experiment: This step involves all ninety-six videos played in a counter-balanced order to each participant to. This is the step where the participant data is collected for analysis.

3.1.1 Simulation. As mentioned at the beginning of Sect. 3, 96 robot trajectories were generated that correspond to different cue combinations, object configurations and objects. These trajectories were generated for a simulated UR5 Robot with a Barrett Hand. ROS [17] and several ROS packages (Gazebo robot simulator [16], MoveIt! [21], RViz [14] and trac_ik [5]) were used to generate simulations as well as ROS-Industrial's [11] Universal Robots UR5 and Robotnik Automation's Barrett Hand BH8-282 [2] descriptions. In the simulation environment, the arm sits on top of a platform, in front of a table that has one red and one blue cylindrical object on top. The camera is placed such that the participant will see the robot from the opposite side of the table to have the same participant-experimenter configuration in the Human-Human Experiment. The resulting setup can be seen in Fig. 4.

3.1.2 Graphical User Interface. The graphical user interface is developed using TkInter Python package with GStreamer as the video player. The GUI shows an information form to be filled out at the beginning of the experiment. Then it goes through the preview, tutorial and actual experiment steps as described in Sect. 3.1. A brief introduction text is provided before each step. Finally, it displays participant statistics at both the tutorial's and actual experiment's end screen, as depicted in Fig. 5b.

3.1.3 Hypotheses. Based on our results of the Human-Human Interaction experiment, we hypothesize that the observed cues can be useful for communicative motion generation by a robot. Furthermore, we think that certain cue combinations can result in different outcomes as compared to being just by themselves. As such, we define our hypotheses as:

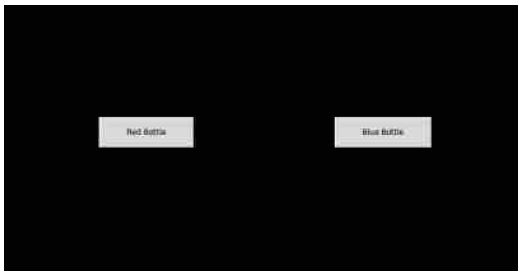
(1) The individual communicative cues (*MA, OT, Exg, GM*) affect the expressiveness of the robot motion compared to the no cue (*Non*) case.
(2) There are certain cues that interact with others to affect the expressiveness of the robot motion.
(3) All the cue combination cases affect the expressiveness of the motion compared to the no cue case.

The first and second hypotheses are evaluated using a multi-factor repeated measures ANOVA with all possible combinations. The third hypothesis is evaluated by paired t-tests between the *Non* case and all other cue cases, for a total of 15 comparisons.

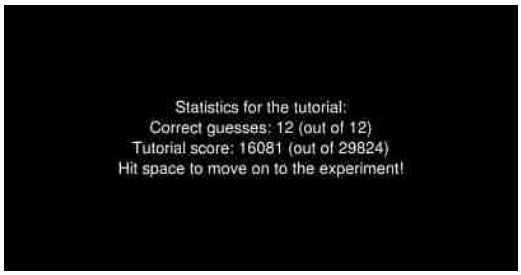
(a) No Intent Expression at t = 0 ms (b) No Intent Expression at t = 767 ms (c) No Intent Expression at t = 1500 ms (d) No Intent Expression at t = 2000 ms

(e) MA + OT + Exg at t = 0 ms (f) MA + OT + Exg at t = 767 ms (g) MA + OT + Exg at t = 1500 ms (h) MA + OT + Exg at t = 2000 ms

Figure 4: Comparison of no intention case vs. maximum aperture + overall time + exaggeration case.



(a) Object selection screen



(b) Result statistics screen

Figure 5: GUI screens at different phases of the experiment.

3.2 Participants

20 male and 10 female participants from a campus community with a median age of 22 attended the experiment. 25 of them stated that they have only seen a robot in videos, 3 of them stated that they have been interested in robotics but did not have any hands-on experience and 2 of them had minor interactions with robots but not in a research related way.

The participants are briefed about the study, GUI usage and the participant score (Eq. 7) before the experiment starts. They are told that the simulated robot does not try to mislead them. To motivate the participants, a gift card of 13$ for a local coffee shop is promised to one of them, selected randomly from the top three.

3.3 Data Collection

The data is collected in the Actual Experiment step. For each video the participant watched, the following are recorded:

(1) Object configuration
(2) Target object robot is reaching towards (*target*)
(3) Active communicative cues (see beginning of Sect. 3)
(4) Time the participant has stopped the video (t_{guess})
(5) Duration of the video (t_{dur})
(6) The participant's object guess (*guess*)

The following measures are calculated for each participant and each video using the collected data:

Correctness (C): Whether the participant's guess of the target object is correct or not. Using an Iverson bracket:

$$C = [guess = target] \tag{4}$$

Guess rating (r_{guess}): The relative point of the video the participant has made the guess. It is a ratio between the time the video was stopped to make a guess and the whole video duration.

$$r_{guess} = \frac{t_{guess}}{t_{dur}} \tag{5}$$

Guess score (J_{guess}): Participant's guess score. It is equal to r_{guess} if $C = 1$ for that video and 1 if $C = 0$ (i.e. it is the same as not being able to make a guess during the whole duration of a video). This score will be used to test the hypotheses.

$$J_{guess} = \begin{cases} r_{guess}, & \text{if } C = 1 \\ 1, & \text{otherwise} \end{cases} \tag{6}$$

Using the variables above, the following two measures are calculated per participant, given n as the number of videos:

Participant score (S): This measure is for motivating the participant to make fast and accurate guesses. This score is inversely proportional with the guess score.

$$S = \sum_{i=1}^{n} (C_i(t_{i,dur} - t_{i,guess})) \tag{7}$$

Table 1: Multi-factor Repeated Measures ANOVA results. Only the factors with p< 0.05 are shown here due to space restrictions.

Source	SS	df	MS	F	p-value
OT	0.7259	1	0.7259	20.3779	<1e-5
Exg	1.0167	1	1.0167	23.0617	<1e-5
GM	4.7189	1	4.7189	101.1136	<1e-6
OT+GM	0.1384	1	0.1384	4.1701	<0.05
Exg+GM	2.0706	1	2.0706	61.5093	<1e-6
Error	109.84	2864	0.04		
Total	118.75	2879			

Number of correct guesses ($n_{correct}$):

$$n_{correct} = \sum_{i=1}^{n} C_i \tag{8}$$

3.4 Experimental Results

3.4.1 ANOVA. As mentioned in Sect. 3.1.3, a multi-factor repeated measures ANOVA with all possible interactions across factors was employed to evaluate hypotheses 1 (H1) and 2 (H2). H1, the cues having effect on the expressiveness of the motion as measured by the guess score (J_{guess}, Eq. 6), is evaluated based on the main effect results of the analysis. H2, unique effects of cue combinations, is evaluated based on the interaction effects of the analysis. The significant results of the ANOVA is presented in Table 1.

H1: The table shows that all the cues but *MA* had a significant effect on the guess score. It can be concluded that this hypothesis is partially satisfied. As discussed in Sect. 4, we think that the cue *MA* is important for intent expression to highlight the action in progress (reach-to-grasp) and not the target of the action, which is what this experiment looked at.

H2: The table shows that there are two cue combinations that result in interaction among cues: *OT+GM* and *Exg+GM*. It can be concluded that this hypothesis is satisfied. It is difficult to interpret this without further analysis which will follow. It should be noted that three and four cue interactions were also included in the analysis but they did not yield any significant results. The lack of significant results for cue combinations do not imply that these combinations have no effect but that the cues do not interact to influence the effect directly.

3.4.2 Paired t-tests and Individual Confidence Intervals. Following the ANOVA, we conduct 15 paired t-tests between the no cue case and all other cue cases, based on the guess score. The previous analysis showed which individual cues and cue combinations had significant effect on the guess score but it did not show whether they had a positive and negative impact, which is needed to evaluate hypotheses 3 (H3). Moreover, the former did not show the effects of all the combinations. With multiple paired t-tests, we can see the effect of individual combinations at the cost of losing statistical power since each comparison will now have one eighth of the data to work with. Furthermore, we calculate the 95% confidence intervals of the differences between the guess score of the no cue case and the other cases, to aid in the analysis. The results of the t-tests are given in Table 2 and the confidence intervals in Fig. 6.

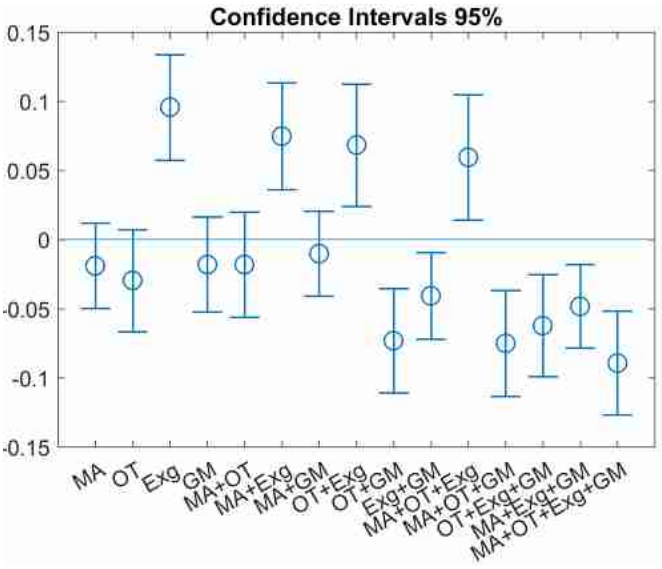


Figure 6: 95% Confidence Intervals for the differences between cue combinations vs. Non case.

H3: 10 out of 15 cases showed significantly better results than the *Non* case based on the table and the figure. *OT* and *GM* were found to have significant effects in the ANOVA study. The discrepancy is not a contradiction but it is due to the lower power of this analysis as explained in Sect. 3.4.2. Combining both analyses, the number of significant cases become 12. However, only 8 out of these 12 cases resulted in a better guess score. The interesting observation is that the 4 negative cases all have *Exg* in common. It can be concluded that this hypothesis is partially and conditionally satisfied.The results are presented without multiple comparison correction in the table and the figure to see the direction of the effects clearly. When we apply Tukey's method to correct for multiple comparisons, only the *Exg+GM* combination loses significance.

As expected, the combination of all four cues result in the best guess score. The surprising result was that exaggeration was not be as good an intent expressing cue as previously suggested. However, this might be due to having only a single view point with bad depth perception (videos viewed on a screen) and only a single direction of exaggeration (away from the non-target object in the horizontal plane). Exaggeration might require a more detailed formulation such as taking the observer view-point and object configurations into account and keeping in mind that it could backfire.

4 DISCUSSION AND FUTURE WORK

Now that cage-free robotics paradigm is near, intent expression and recognition should be investigated to improve fluency and efficiency of human-robot collaboration. In this study, reach-to-grasp arm motions were investigated to identify subtle communicative cues in non-verbal interactions between humans. Our hypothesis was that subtle non-verbal cues of human communicative motion can be utilized in human-robot interaction and collaboration.

The human-human experiment (Sect. 2) helped us identify four subtle communicative cues: (1) An increase in overall motion duration, (2) a decrease in the relative time to reach maximum hand

Table 2: Mean guess scores along with significance results for the paired tests. Red box is the worst and green box is the best guess score. Degrees of freedom is 358 for all.

Cue Combinations	Mean	t	p-values
No Intention	0.5334	N/A	N/A
MA	0.5145	1.0956	>0.05
OT	0.5036	1.5884	>0.05
Exg	0.6290	-4.6423	<1e-4
GM	0.5154	1.0203	>0.05
MA + OT	0.5152	0.8661	>0.05
MA + Exg	0.6081	-3.6584	<1e-3
MA + GM	0.5233	0.5727	>0.05
OT + Exg	0.6016	-2.9788	<1e-2
OT + GM	0.4603	3.8245	<1e-3
Exp + GM	0.4927	2.4692	<0.05
MA + OT + Exg	0.5928	-2.6275	<1e-2
MA + OT + GM	0.4584	3.7746	<1e-3
OT + Exg + GM	0.4712	3.3094	<1e-3
MA + Exg + GM	0.4851	2.9786	<1e-3
MA + OT + Exg + GM	0.4441	4.6579	<1e-4

aperture, (3) exaggeration of motion, and (4) grasp modality change between configurations and between objects. Unexpected participant responses to survey questions showed two other interesting finds. Firstly, none of the participants were aware of their change about relative time to maximum hand aperture. This means, hand aperture timing might be one of the key aspects of non-verbal intent expression for reach-to-grasp motions. Secondly, some participant responses showed that the whole body posture might change during reach-to-grasp motions to express intent.

The answers to survey questions show that there are unidentified cues yet to find in even this simple task. The answers also suggest that human-human interaction should not simply be observed but participants should be asked about their views on the study and what seems to be important to them. These answers might help with identification of other natural and non-verbal communication cues and features in human-human scenarios. Participants' responses seem to demonstrate that these elements would not only help with motion generation towards intent expression but would also help with recognizing motion intent. If these elements are learned by a robot, the robot could both generate motion and recognize human-motion intent. This is an interesting avenue for future work, e.g., tracking human torso orientation wih 3D cameras for intent recognition in human-robot collaboration.

Following the human-human experiment, we have conducted a simulated robot experiment (Sect. 3) to understand which of the identified subtle motion cues help in generating intent expressive robot motion. The participants used a GUI which showed videos of a simulated robot reaching to grasp an object among two and they were asked to identify the target as early and accurately as possible. This experiment had a complex multi-factorial design which required generation of trajectories based on all cue combinations. The results showed that overall time duration, grasp modality and certain cue combinations result in better guesses than no cue condition. Relative time to reach maximum hand aperture had no effect on the guess quality. We think that this cue is actually about expressing the action being done rather than about the target of the action. An interesting future work is to validate this claim by comparing it to other actions such as reach-to-touch. Exaggeration had a negative effect on the guess quality, unless combined other cues. We think that this cue needs a more complicated formulation based on observer view-point and object configurations and may not be suitable for videos.

In this study, we did not test the identified communicative cues in actual human-robot scenarios. The reason is that the two experiments allowed us to identify and systematically test the cues before a third HRI experiment. We think that application of such communicative cues in a real-world robotic scenario would increase fluency and efficiency of human-robot collaboration. An immediate next step is to extend our work towards this direction.

This work did not consider scenarios other than reach-to-grasp motions, and even in this scenario, only considered cylindrical objects. Avenues of extension include considering different objects (e.g. paper or hammer), reaching into clutter or a task-based scenario such as collaborative assembly.

The subtle cues we have identified can be used as features in a machine learning approach, along with other information such as whole-body pose, gaze and object locations and object features for intent expression. Several machine learning approaches are possible. One example is inverse reinforcement learning or inverse optimal control to learn rewards/costs and generating motion through planning or trajectory optimization. Another approach could be learning spatio-temporal constraints (e.g. for motion duration or when to reach maximum hand-aperture) to be used in planning. Learning the values of the identified cues through policy gradient methods for motion generation [7] and directly learning motion models, for example with interaction primitives, [1] is also viable. These approaches would automate motion generation and some of them can be extended to intent recognition as well.

5 CONCLUSION

This paper presented two experiments. The first one, a human-human reaching-to-grasp experiment, identified subtle cues to be used in robot motion generation. The second one, a simulated robot experiment, showed that these cues and their combinations can be used to generate intent expressive robot motion with a surprising exception. To the best of our knowledge, this is the first study that looks at humans to get intent expressive cues in the context of arm motions and that tests these cues in a systematic study with a complex multi-factorial experiment design. Through this study, it was shown that changes solely in spatial aspects of motion might not be enough for intent expression in HRI. This study also shows that further investigation in human-human scenarios might be needed to identify spatio-temporal cues in different tasks to automate intent expressive motion generation for improving fluency and efficiency in human-robot collaboration.

There are interesting future directions such as conducting an experiment with a real robot, testing these cues in more realistic scenarios, evaluating cues for expressing the action in addition to its target and automating intent expressive motion generation.

REFERENCES

[1] Heni Ben Amor, Gerhard Neumann, Sanket Kamthe, Oliver Kroemer, and Jan Peters. 2014. Interaction primitives for human-robot cooperation tasks. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*. IEEE, 2831–2837.

[2] Robotnik Automation. 2014. Barrett Hand Descriptions. (2014). Retrieved Nov. 5, 2017 from https://github.com/RobotnikAutomation/barrett_hand_common

[3] Cristina Becchio, Valeria Manera, Luisa Sartori, Andrea Cavallo, and Umberto Castiello. 2012. Grasping intentions: from thought experiments to empirical evidence. *Frontiers in Human Neuroscience* 6 (2012), 117. https://doi.org/10.3389/fnhum.2012.00117

[4] Cristina Becchio, Luisa Sartori, Maria Bulgheroni, and Umberto Castiello. 2008. The case of Dr. Jekyll and Mr. Hyde: A kinematic study on social intention. *Consciousness and Cognition* 17, 3 (2008), 557 – 564. https://doi.org/10.1016/j.concog.2007.03.003

[5] Patrick Beeson and Barrett Ames. 2015. TRAC-IK Kinematics Plugin. (2015). Retrieved Nov. 5, 2017 from http://wiki.ros.org/trac_ik

[6] Christopher Bodden, Daniel Rakita, Bilge Mutlu, and Michael Gleicher. 2016. Evaluating Intent-Expressive Robot Arm Motion. In *International Symposium on Robot and Human Interactive Communication*. IEEE.

[7] Marc Peter Deisenroth, Gerhard Neumann, and Jan Peters. 2013. A Survey on Policy Search for Robotics. *Found. Trends Robot* 2, 1–2 (Aug. 2013), 1–142. https://doi.org/10.1561/2300000021

[8] Anca Dragan, Shira Bauman, Jodi Forlizzi, and Siddhartha Srinivasa. 2015. Effects of Robot Motion on Human-Robot Collaboration. In *Human-Robot Interaction*. Pittsburgh, PA.

[9] Anca Dragan, Kenton Lee, and Siddhartha Srinivasa. 2013. Legibility and Predictability of Robot Motion. In *Human-Robot Interaction*. Pittsburgh, PA.

[10] Anca Dragan and Siddhartha Srinivasa. 2013. Generating Legible Motion. In *Robotics: Science and Systems*. Pittsburgh, PA.

[11] Shaun Edwards and Chris Lewis. 2012. Ros-industrial: applying the robot operating system (ros) to industrial applications. In *IEEE Int. Conference on Robotics and Automation, ECHORD Workshop*.

[12] F. Faul, E. Erdfelder, A. Buchner, and A. G Lang. 2009. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods* 41 (2009), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

[13] M. J. Gielniak and A. L. Thomaz. 2011. Spatiotemporal correspondence as a metric for human-like robot motion. In *2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 77–84. https://doi.org/10.1145/1957656.1957676

[14] Dave Hershberger, David Gossow, and Josh Faust. 2008. RViz: 3D visualization tool for ROS. (2008). Retrieved Nov. 5, 2017 from http://wiki.ros.org/rviz

[15] Ollie Johnston and Frank Thomas. 1995. *The Illusion of Life: Disney Animation*. Hyperion.

[16] Nathan Koenig and Andrew Howard. 2004. Design and use paradigms for gazebo, an open-source multi-robot simulator. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings.*, Vol. 3. IEEE, 2149–2154.

[17] Morgan Quigley, Ken Conley, Brian P. Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y. Ng. 2009. ROS: an open-source Robot Operating System. In *ICRA Workshop on Open Source Software*.

[18] Luisa Sartori, Cristina Becchio, Bruno G. Bara, and Umberto Castiello. 2009. Does the intention to communicate affect action kinematics? *Consciousness and Cognition* 18, 3 (2009), 766 – 772. https://doi.org/10.1016/j.concog.2009.06.004

[19] Luisa Sartori, Cristina Becchio, and Umberto Castiello. 2011. Cues to intention: The role of movement information. *Cognition* 119, 2 (2011), 242 – 252. https://doi.org/10.1016/j.cognition.2011.01.014

[20] Natalie Sebanz, Harold Bekkering, and Günther Knoblich. 2006. Joint action: bodies and minds moving together. *Trends in cognitive sciences* 10, 2 (2006), 70–76.

[21] Ioan A. Sucan and Sachin Chitta. 2013. MoveIt! (2013). http://moveit.ros.org

[22] Min Zhao, Rahul Shome, Isaac Yochelson, Kostas Bekris, and Eileen Kowler. 2016. *An Experimental Study for Identifying Features of Legible Manipulator Paths*. Springer International Publishing, Cham, 639–653. https://doi.org/10.1007/978-3-319-23778-7_42