To appear in IEEE International Symposium on Robot and Human Interactive Communication 2015

Visual Pointing Gestures for Bi-directional Human Robot Interaction in a Pick-and-Place Task

Camilo Perez Quintero, Romeo Tatsambon, Mona Gridseth, Martin Jägersand*

***Abstract*— This paper explores visual pointing gestures for two-way nonverbal communication for interacting with a robot arm. Such non-verbal instruction is common when humans communicate spatial directions and actions while collaboratively performing manipulation tasks. Using 3D RGBD we compare human-human and human-robot interaction for solving a pick-and-place task. In the human-human interaction we study both pointing and other types of gestures, performed by humans in a collaborative task. For the human-robot interaction we design a system that allows the user to interact with a 7DOF robot arm using gestures for selecting, picking and dropping objects at different locations. Bi-directional confirmation gestures allow the robot (or human) to verify that the right object is selected. We perform experiments where 8 human subjects collaborate with the robot to manipulate ordinary household objects on a tabletop. Without confirmation feedback selection accuracy was 70-90% for both humans and the robot. With feedback through confirmation gestures both humans and our vision-robotic system could perform the task accurately every time (100%). Finally to illustrate our gesture interface in a real application, we let a human instruct our robot to make a pizza by selecting different ingredients.**

I. Introduction

Robot arm manipulation in household robotics has been studied for more than two decades, yet robots are still not capable of dealing with home environments [1]. Household environments pose a challenge for robots because they are unstructured and dynamic. Recently, several researchers have been working towards solving this challenge [2]. Robot automation has been useful for industry, but robots in household situations require a different interface. A better approach consists of robots capable of collaboratively interacting with humans [3], [4]. Non-verbal gesture communication is a powerful human interaction. Lozano and Tversky [5] conducted experiments where human participants performed an assembly task instructed by speech or gestures. Participants understood and learned better from gesture-only than from speech-only instructions. Important human-robot interaction cues can be learned from human-human non-verbal communication.

Our work focuses on human-robot non-verbal communication, where human-robot interaction is performed without devices such as tablets and control panels. Instead gestures are used to let communication occur directly between humans and

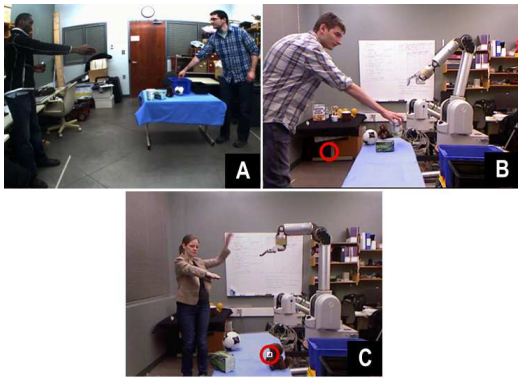

Fig. 1: Instructor and assistant interaction
A. Human-Human: Human instructor selects an object by pointing. A human assistant asks if his interpretation is correct.
B. Robot-Human: Robot instructor selects an object by pointing. A human assistant interprets the pointing gesture performed by the robot, approaches the table and picks the selected object.
C. Human-Robot: Human instructor selects an object by pointing. A robot assistant is ready to reach the selected object.

*This work is supported by iCORE, NSERC and the Canadian Space Agency (CSA).

Authors are with the Department of Computing Science, University of Alberta, Edmonton, AB, T6G2E8, Canada. caperez@cs.ualberta.ca

robots. This paper aims at studying and developing communication mechanisms that allow humans to intuitively instruct or cooperatively achieve a manipulation task with a robot, that uses a gesture-based human-robot interface (HRI). In contrast to the other works mentioned above, we will focus on the visual interpretation of pointing in a pick-and-place task (Section II) and the interaction between the human and the robot for solving the task. Pointing to indicate direction or position is one of the intuitive communication mechanisms used by humans in all life stages. However, misinterpreting a pointing gesture could lead to a wrong direction or position. Humans have the capacity to corroborate gesture interpretations by interacting with each other. An immediate goal of this paper is to give the robot the same functionality during human-robot interaction. We think that researching human-like interfaces will bring robots closer to becoming useful in home environments. Our contributions are:

- A human-robot gesture language for two-way communication in pick-and-place tasks.
- A vision system using a Kinect that can: detect gestures, detect objects location on an horizontal plane, detect human pointing direction and infer a 3D pointing location in the scene. Based on that location the system can infer what object is being pointed to and return a possible grasping location. If the return object is incorrect the system is capable of interaction with the human until the right object is found.
- A behavioural state machine that implements the pick-and-place task interaction.
- An experimental performance and error evaluation study where 8 human subjects use the robotic system either as an instructor or assistant in a collaborative task.
- A practical application of our system, where our robot prepares a customized pizza by interacting with a human by gesturing.

II. Task

Experiments are done with a pick-and-place task example application. The objective is to clear objects off a table and sorting the objects in the appropriate containers. In the task we have two actors: instructor and assistant, and two types of communication: non-feedback and feedback. The actions performed by the actors for both types of communication are described in the work-flow diagram shown in Figure 3.

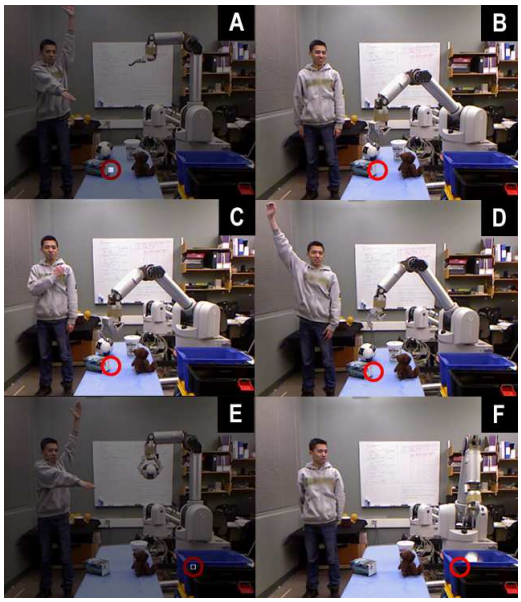

Fig. 2: A human performing a sequence of gestures to interact with the robot
A. Human instructor selects a desired object by pointing. **B.** Robot assistant interprets the pointing gesture performed by the human and performs a confirmation gesture to the human instructor. **C.** Human instructor denies robot interpretation by crossing his dominant arm over his torso. **D.** Robot assistant moves to the next possible selected object. Human instructor confirms the robot interpretation by raising his dominant hand. **E.** Robot assistant picks up the selected object. Human instructor selects a desired dropping location **F.** Robot assistant places the object in the selected container.

We are covering human-human, robot-human and human-robot interaction (see Figure 1). Our system in operation for the human-robot case is shown in Figure 2 (the red ring indicates the pointing location detected from the depth image projected into the RGB image). For the non-feedback steps B, C and D (see Figure 2) are removed. Our research focuses on pointing gestures because of their simplicity and universal understanding. However, to enable complete interaction, we include Yes and No symbolic gestures, both for robot-human and human-robot interaction.

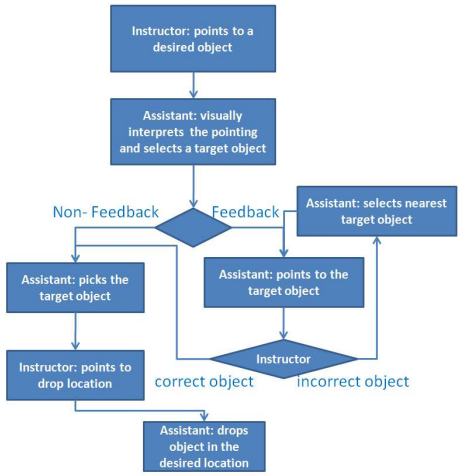

Fig. 3: Pick-and-place task work-flow diagram for both feedback and non-feedback communication cases.

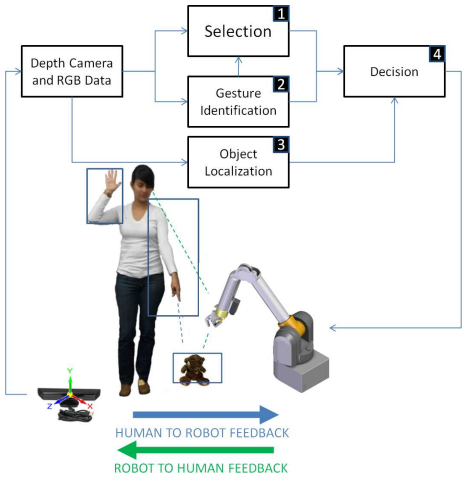

Fig. 4: System Diagram. Through the human gesture identification, object localization and pointing interpretation:
*The robot gets feedback from the user by detecting human gestures (blue arrow) and interpreting it.
*The human gets feedback from the robot by interpreting gestures performed by the robot (green arrow).

The way the symbolic gestures are expressed is not important for the pointing evaluation here. They could, for instance, be replaced by verbal Yes and No. A demonstration of our system can be seen in the accompanying video or in: http://webdocs.cs.ualberta.ca/%7Evis/HRI/PickandPlace.wmv

III. System Description

Our system provides a two-way communication channel based on gestures between a human and a robot to achieve a pick-and-place task. Below is a description of our human-robot and robot-human system. For both interactions our system uses a 7DOF WAM arm, a Microsoft Kinect sensor and a regular Linux machine.

A. Human-Robot System

Our human-robot system is illustrated in Figure 4. We assume that the human and the objects are in the field of view of our depth camera and that the possible selected objects are located in a plane of interest that the robot can reach. Figure 5 shows the point cloud and RGB visualizations of our system respectively. The human does not see the visualizations, but interacts solely with the physical robot. In Figure 5A, the user points to a desired object. Notice the virtual red line that is generated using the red sphere inside the user's head and the teal sphere inside the users hand. After the hit location is found (red sphere at the end of the virtual ray) the system corrects the hit location to the nearest object and proposes a possible grasping location on top of the object (dark blue sphere). Figure 5C shows the identified objects in the system (green spheres) and the grasping location (dark blue sphere). The system gets ready to give feedback to the user.

Our system is composed of four blocks: Selection, Gesture Identification, Object Localization and Decision (Figure 4).

1) Selection: The Selection block is based on our implementation [6], where the user points to a target object or location and the interface returns the 3D hit position coordinates. The input for the Selection block is the depth sensor data, and the output is the (x,y,z) hit location. Based on the depth information, the user's head and hand position are found and used to calculate the pointing direction. Notice that in [6] we focused on the technical accuracy analysis of human pointing direction only, and we neither completed a task nor used a robot. In contrast, in this work we are interactively solving a task with a robot using pointing gestures.

2) Gesture Identification: Based on the pick-and-place task explain in Section II, we define four human body gestures for the human-robot interaction. These are **Stand By** (Figure 2B), **Pointing** (Figure 2A), **Yes** (Figure 2D) and **No** (Figure 2C). Our Gesture Identification block is in charge of interpreting the predefined human gestures. We use the OpenNI skeleton tracking libraries to find human joint locations. Gesture identification is based on spatial relations between the different human joint locations.

3) Object Localization: This block receives data from the depth camera as input, and outputs the centroid locations and bounding box of one or more objects located on the plane of interest (Figure 5C). We use the point cloud library (PCL) to manipulate data coming from the depth sensor.

The point cloud obtained from the Kinect is downsampled with a voxelized grid approach. Then using RANSAC [7] and a 2D convex hull, we find the table plane coefficients, inliers and points that belong to the table. Next, the inliers are clustered by distance to distinguish the objects. Finally, the mean vector for each cluster and its minimum bounding box are calculated.

4) Decision: The Decision block provides our system with the capacity of interaction. This block outputs the robot interaction and receives as input the hit point location from the Selection block, the identified gesture preformed by the human from the Gesture Identification block and the object centroids and bounding boxes from the Object Localization block (See Figure 4). The Decision block is based on the state machine shown in Figure 6, which consists of six states. The human is interacting using the **Stand by** gesture and the system holds until a **Pointing** gesture is performed by the human (State 1). Then the system saves the hit point coordinates coming from the Selection block (State 2). The system holds until the human goes back to the **Stand by** gesture. Using the input data from the Object Localization and Selection blocks, the system calculates the nearest object centroid to the hit point (State 3). The system projects the centroid position into the top face of the object bounding box (dark blue sphere in Figure 5).

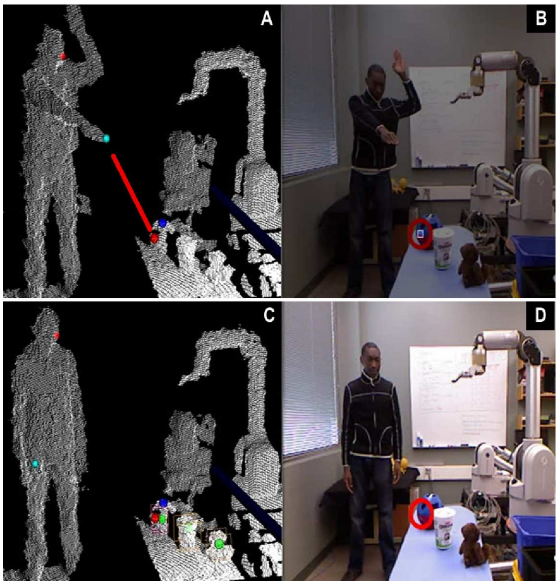

Fig. 5: System point cloud visualization (A,C) and RGB visualization (B, D). Centroids (green spheres) and bounding boxes extraction from objects over the table plane (C). The system corrects the the ray hit point in the scene (red sphere) to the closest detected object (shape sorter toy).

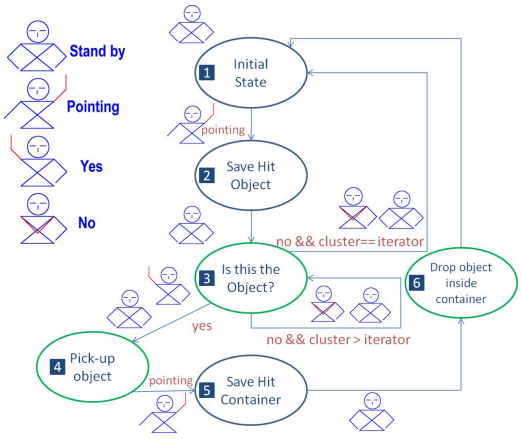

Fig. 6: Finite State Diagram. The system is activated when the user selects an object by pointing. Six states are needed for the complete interaction.

The robot, which has been previously calibrated with respect to the depth sensor camera, uses the centroid projection location to locate the robot end effector over the selected object and gives feedback to the human by performing a pointing gesture (see Figure 2B). The system holds until the human performs a **Yes** or **No** gesture. If the gesture is **No** followed by a **Stand by** gesture, the system iterates to the next nearest object and returns to state 3. In the case where the human keeps doing the **No** gesture until there are no objects left the robot goes back to state 1. If the the human performs a **Yes** gesture followed by a **Stand by** gesture the state is shifted to state number 4. Using the selected object centroid projection (dark blue sphere in Figure 5), the robot locates its hand above the projected object centroid with its palm perpendicular to the table plane and grasps the object. Then the robot goes to an initial position and waits for a **Pointing** gesture indicating the dropping location (State 4). After the **Pointing** gesture is performed the hit target is saved and compared with the possible available container locations (State 5). The system chooses the container closest to the hit point. The robot drops the object in the container and goes back to the initial position (State 6). In our system the decision block is tailored to the pick-and-place example application. However, our system is based on a general state machine representation and as shown by [8] a large variety of tasks for home robots can be implemented with such a system. In fact in our application example we do that.

B. Robot-Human system

In the robot-human case the gesture interpretation is done by the human (see Figure 1B), which makes the system implementation simpler than in the human-robot case. We define four gestures for the robot: **robot-Stand-by**, **robot-Pointing** (Figure 1B), **robot-Yes**: The robot moves its right finger up and down repeatedly, **robot-No**: The robot moves its wrist from right to left repeatedly. In contrast with our human-robot configuration, where objects are dynamically detected we used predefined positions for implementation simplicity. We pre-record the robot pointing gesture directions for the different object locations in the two test configurations; objects on a line and general (objects spread out over the table). We interact with the user by predefined robot configurations.

IV. Experiments and Analysis

We performed three experiments using the instructor-assistant pick-and-place task described in Section II, where the actors for each experiment are: human-human, human-robot and robot-human. We had a total of 8 participants aged from 18 to 34 with 6 male and 2 female. Among them 5 had corrected vision and one was left handed. The average time per participant to complete the three experiments was 1 hour including break times.

A. Experimental setup

The different experimental setups are shown in Figure 1. Two arrangements of objects were used: general and line. In the general configuration objects are spread over the table surface. In the line configuration, objects are collinear with the instructor line of sight. In the second case only the arm tilt angle is informative, while in the former case, both tilt and pan angles help indicate what object the instructor points to. For both robot-human and human-robot experiments the 7DOF WAM arm was located such that the table space belonged to the robot arm workspace (see Figure 1B and Figure 1C).

B. Human-human

The human-human case (Figure 1A) took an average of 3.05 min per participant. In the first set of experiments we considered a general configuration of 4 objects on the table. In the second set we considered differently ordered line arrangement of the 4 objects (Figure 5B). The line arrangement is purposely made to be aligned with the pointing direction of the instructor. The motivation of using this configuration is to test experimentally whether when a human is inferring a pointing direction, the point of view matters. The feedback experiments took the longest (1.1 min) since participants needed more time to use the pre-determined gestures (Section III-A.2), which have been introduced to them at this stage. Experimental results are shown in

Figure 8. In the human-human case success ratio for the line configuration is 0.75 while for the general configuration is 0.95. It is clear then that it is harder to interpret the pointing in the former case without any feedback. In fact, the pointing with line arrangement of objects has been designated by the participants as the most difficult pointing configuration to interpret.

C. Robot-human

The robot-human case (Figure 2) took an average of 8.15 min. Here we performed a set of 3 experiments with 8 participants. We considered both the general arrangement configuration and the line arrangement configuration of the objects. Results are shown in Figure 8, where we can see that the general object configuration without feedback is equally difficult as the line configuration without feedback. We believe that the difference in results obtained between the human-human case and the robot-human is due to the under actuation in the Barrett hand Barrett hand, which makes it difficult to interpret pointing location.

D. Human-Robot

The human-robot case (Figure 2C) took on average 2 min. Here we investigate how our system interprets human pointing as well as the accuracy and precision of our system. Accuracy is defined as the mean value of the sample data, whereas precision is the standard deviation, *i.e.,* Euclidean distance between hit point (red sphere) and the object location (green sphere), and precision as the uncertainty (Figure 5C). The average experimental accuracy obtained during the experiments was 8.15 cm and precision 0.17 cm. This means that our Decision block has to deal with this uncertainty and correct it in order to obtain a suitable grasping location. In the human-robot interaction case the success ratio is equal in both the general configuration and the line configuration, see Figure 8. This means that our vision system can better interpret human pointing than a human assistant.

Figure 7 shows that the number of misinterpretations of the assistant is lowest in the human-robot interaction. In fact there is 28% misinterpretations

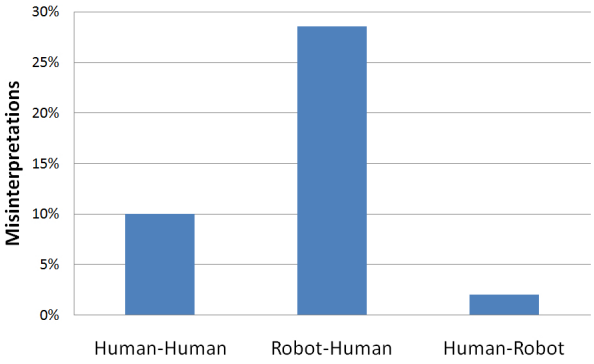

Fig. 7: Percentage of feedback misinterpretations

in a robot-human interaction, 10% misinterpretation in a human-human interaction and 2% misinterpretation in a human-robot interaction. This result means that our system is better in this particular case at interpreting pointing than humans.

The high percentage of misinterpretations in the human-robot case might be explained by the Barrett hand being non-anthropomorphic and not able to point with a straight finger. This is a technical limitation with this hand because of the underactuation constraints used to build it.

On a Likert-type scale from 1 to 7, in average participants find it equally difficult, less than 2, to accomplish the task with feedback in the human-human, robot-human and human-robot interactions. Understanding feedback was equally difficult in the three interaction cases. Furthermore, participants find it equally difficult, around 4, to interpret both human pointing and robot pointing in the robot-human interaction. The reason in the first case is obvious since it is generally difficult for another human to understand human pointing. The reason in the second case is that participants were confused with the robot pointing because the robot finger was not straight. Human-human pre-determined gestures for communication, interpretation of robot gestures, use of human pre-determined gestures and pointing to the right object, were found to be not so difficult. Finally human-human gesture interpretation was found less difficult to interpret as humans are used to interact with other humans. To sum up, our experiments showed that neither

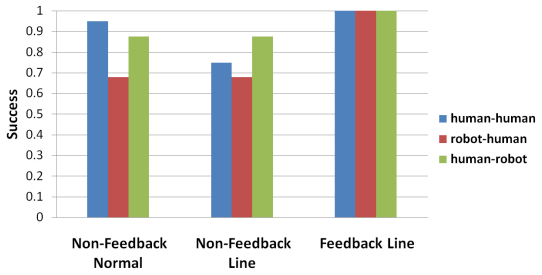

Fig. 8: Commander and assistant experiment success rate

humans nor our robot vision system could interpret pointing perfectly (Though the robot vision system was more accurate than humans). Therefore the extended interaction through confirmation gestures we introduced is important to make HRI reliable.

E. Other gestures used in human-human interaction

During the human-human experiments we told the subjects to non-verbally communicate what items they wanted the other person to pick up, but we did not tell them what specific gestures to use. This allowed us to observe what gestures the subjects chose. Most of the subjects simply pointed towards the target object. However we also saw some other variations. A couple of subjects used their hands to indicate the shape of the desired objects. That is, they formed a round shape to show they wanted the soccer ball. Similarly another subject used this same technique to tell the other person to pick up the yogurt container. Another gesture we came across was a kind of counting gesture. When the objects were ordered in a line, the subject moved his/her forearm in a circular fashion three times to indicate he/she wanted the third object. We also observed another technique to distinguish between objects at different distance from the subject. Two subjects stood on their toes while pointing toward target objects that were farther away. In particular this was also used to distinguish between the blue and the black containers. Finally one of the subjects used curved pointing when the objects were placed in a line. Instead of pointing directly towards the object in a straight line, the subject formed a curve with his/her arm to allow for more precise selection. These observations give us new ideas for improving our current pre-determined gestures.

V. Application example: "Making pizza with my robot"

We envision that our system can be used in different real life scenarios, e.g., a robot can work behind a counter taking the role of a shopkeeper; a client points to a particular object and by using confirmation feedback the robot will reach the desired product. In another situation a robot can be used as a chef at a hotel breakfast buffet; the client points to different ingredients to include in his omelette. In a metal workshop a robot can assist a welder by picking and placing parts. The welder only has to point to them, avoiding heavy weight manipulation and extreme temperatures. To bring our study to a practical situation we made our robot capable of preparing pizza by gesturing with a human. The application set-up is shown in Figure 9. Ingredients are randomly placed on top of the table and detected. When the user gets close to the cooking table the robot is activated and the human tracking starts. The user can then select any ingredient by simply pointing to it. After the selection the robot picks the ingredient and pours it on top of the pizza tray, see Figure 10 A-C. This action can be repeated as many times as the user wants. After the user is satisfied with the number and amount of ingredients, a "finish pizza" gesture is performed and the robot places the pizza in the oven, see Figure 10 D-F. A video of the complete interaction can be found at http://webdocs.cs.ualberta.ca/%7Evis/HRI/makingPizza.wmv. The example application demonstrates that a complex task for a robot, like preparing a customized pizza for a client, can be simplified by using the appropriate communication interface.

VI. Conclusions

When humans collaborate on manipulation tasks, gestures form an integral part of the communication. It is often easier to point to an object or desired location than to describe it in words or numbers. We designed and evaluated a robot and vision system

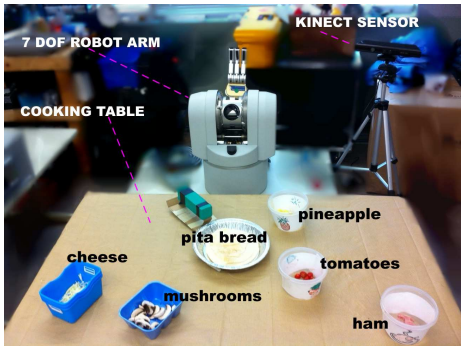

Fig. 9: The set-up used in our practical application "making pizza with my robot" consists of a 7DOF robot arm, a kinect sensor and a cooking table with ingredients.

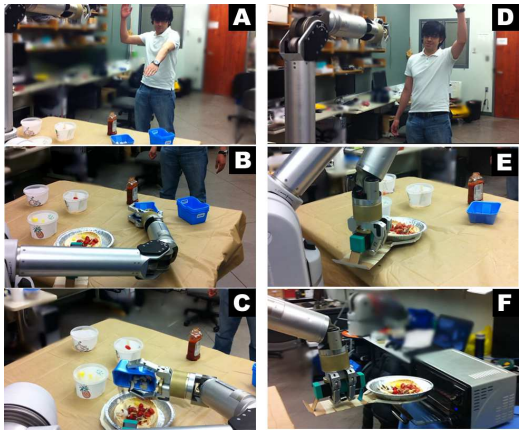

Fig. 10: Left Column: User selects mushrooms by pointing, the robot picks and pours mushrooms in the pizza tray. Right Column: User performs the "finish pizza gesture" and the robot places the pizza inside the toaster oven.

that is able to see, interpret and act using a gesture language. In our experimental study 8 humans interacted with the robot for about 1h each. We prove experimentally that our system can behave similar to and for specific cases better than a human interpreting human pointing. The task was to clean off a table and sort the objects into containers. We compared human-robot, robot-human and human-human pairs as instructors and assistants respectively. We performed the tasks both with instruction only (one way communication) and with feedback gestures from the assistant (robot or human) to verify that the robot/human had interpreted the task correctly. Without feedback the assistant could interpret the pointing gesture correctly in 70-95% of the cases. Humans had particular difficulty distinguishing between objects placed along a line (75% success rate), but were much better with the objects in a general configuration (95% success). Humans also had more difficulty interpreting the robot's pointing (70%) than another human's pointing. This is likely due to the physical inability of our Barrett robot hand to extend the finger fully and point with a straight gesture towards the object. The robot vision system had similar accuracy independent of object configuration (88% success). In the feedback case the assistant indicated the object to select by pointing just above it. The instructor could then confirm with a yes gesture, or deny with a no gesture, and then point again to the desired object. This feedback allowed successful task completion in all cases for both the robot and human assistants. In questionnaire answers the human subjects indicated that for this task the human-robot system was about equally easy to work with compared to human-human communication. Finally we implemented the application "Making pizza with my robot", where we showed how our study can be brought to a practical human scenario. We strongly believe that by researching simple and novel ways of human robot communication will bring robotics closer to human environments.

REFERENCES

[1] C. C. Kemp, A. Edsinger, and E. Torres-Jara, "Challenges for robot manipulation in human environments," *Robotics & Automation Magazine, IEEE*, vol. 14, no. 1, pp. 20–29, 2007.

[2] R. B. Rusu, Z. C. Marton, N. Blodow, M. Dolha, and M. Beetz, "Towards 3d point cloud based object maps for household environments," *Robotics and Autonomous Systems*, vol. 56, no. 11, pp. 927–941, 2008.

[3] S. Satake, T. Kanda, D. F. Glas, M. Imai, H. Ishiguro, and N. Hagita, "How to approach humans?-strategies for social robots to initiate interaction," in *Human-Robot Interaction (HRI), 2009 4th ACM/IEEE International Conference on*. IEEE, 2009, pp. 109–116.

[4] A. D. Dragan, A. L. Thomaz, and S. S. Srinivasa, "Collaborative manipulation: new challenges for robotics and hri," in *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*. IEEE Press, 2013, pp. 435–436.

[5] S. C. Lozano and B. Tversky, "Communicative gestures facilitate problem solving for both communicators and recipients," *Journal of Memory and Language*, vol. 55, no. 1, pp. 47–63, 2006.

[6] C. Perez Quintero, R. Tatsambon Fomena, A. Shademan, N. Wolleb, T. Dick, and M. Jagersand, "Sepo: Selecting by pointing as an intuitive human-robot command interface," in *Robotics and Automation(ICRA)*, 2013.

[7] M. Fischler and R. Bolles, "Random sample consensus," *Communications of the ACM*, vol. 24, no. 6, 1981.

[8] H. Nguyen, M. Ciocarlie, K. Hsiao, and C. C. Kemp, "Ros commander (rosco): Behavior creation for home robots," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*. IEEE, 2013, pp. 467–474.