Simplified Industrial Robot Programming: Effects of Errors on Multimodal Interaction in WoZ experiment

Zdeněk Materna, Michal Kapinus, Michal Španěl, Vítězslav Beran, and Pavel Smrž

***Abstract*— This paper presents results of an exploratory study comparing various modalities employed in an industrial-like robot-human shared workplace. Experiments involved 39 participants who used a touch table, a touch display, hand gestures, a 6D pointing device, and a robot arm to show the robot how to assemble a simple product. To rule out a potential dependence of results on the number of misrecognized actions (resulting, e.g., from unreliable gesture recognition), a controlled amount of interaction errors was introduced. A Wizard-of-Oz setting with three user groups differing in the amount of simulated recognition errors helped us to show that hand gestures and 6D pointing are the fastest modalities that are also generally preferred by users for setting parameters of certain robot operations.**

I. Introduction

Industrial robots were traditionally used mainly in a large-scale production. This was primarily due to the large price of the automation and low flexibility requiring long and costly adaptation for new products. Recently, EU-supported projects as SMErobotics[1] and EuRoC[2] emerged to support development of easily reconfigurable cognitive robots able to achieve flexibility required for small to medium scale manufacturing. Such flexibility must be supported by easy to use and effective human-robot interaction substituting traditional ways of programming industrial robots requiring expert-level knowledge.

Our long-term goal is to create a shared-space environment similar to the experimental setup shown in Figure 1 where a human operator can cooperate with a semi-autonomous cognitive robot using multi-modal interaction and augmented reality: ARTable. The robot within the envisioned solution could be programmed once and then perform independently or it may continuously provide assistance to the operator. There was a research on what modalities are appropriate for what most common operations [1] in such a system. As a first step towards ARTable we were interested in how various modalities would perform in a similar experiment however under realistic conditions. Therefore we designed a WoZ experiment where input modalities were not always working perfectly and participants had to face interaction errors. The aim of the experiment was to uncover whether there is dependence between preference for using particular modality for setting particular parameter and amount of experienced interaction errors. Secondarily, we were interested in how task completion times will be influenced by used modality and amount of errors as a time-effective human-robot interaction will be of paramount importance for a practical usage of such system. Video summary of the experiment can be seen at https://youtu.be/LtiDc3pGjug.

All authors are affiliated with the Brno University of Technology, Faculty of Information Technology, Centre of Excellence IT4Innovations, Bozetechova 1/2, Brno, 612 66, Czech Republic. Contacts: *imaterna, ikapinus, spanel, beranv, smrz@fit.vutbr.cz*

[1]http://www.smerobotics.org

[2]http://www.euroc-project.eu/index.php?id=challenge_1

Fig. 1: Prototype of the human-robot shared-space environment with augmented reality user interface (image edited).

II. Related Work

Robot manipulators used to be programmed by experts at a low level making them less flexible to production changes. Recently, approaches allowing high-level programming by end users appeared. One of these approaches is programming by demonstration [2] also referred to as kinesthetic teaching [3], where an operator programs a robot by positioning its end-effector while learning poses [4] and/or forces [5]. Existing solutions can be divided into those allowing so called offline programming where a robot is programmed once [6], [7], those allowing a continuous human-robot collaboration [8] and those allowing both [9] modes. The interface may be for instance projected [10] or integrated into a hand-held device with augmented reality [6], [7]. Interaction also may happen in a virtual reality [9]. Alternatively to positioning a robot's end-effector, a human operator may demonstrate the task by actually performing it [11] or by giving high-level instructions using one [8] or more modalities [6].

Errors in interaction can be according to [12] divided into following types: misunderstandings, non-understandings and

misconceptions. For our experiment, we choose to simulate misunderstandings with third-turn repair of the errors. Dealing with errors is often limited to resolving problems during program execution [13]. The experiment with social robot programming [14] where gesture and speech-based interfaces and even the robot's software were not perfectly reliable has shown importance of the provided feedback. However, those errors were not simulated and thus their amount was not controllable. The framework to support WoZ studies from [15] allows to insert given amount of random misrecognition errors, however it is limited to the speech-based interfaces.

Misunderstandings may be caused by a non-perfect input. For instance the pointed object estimation from [16] is reported to have 83% success rate despite usage of a prior information about location of the objects. Another approach to detection of pointing directions [17] achieved $\pm 10^{\circ}$ angular and 93% distance error. The speech recognition system from [18] achieved 16% error in a noisy environment with background TV or radio. It can be speculated that amount of errors would be higher in an industrial environment.

III. User Study Design

The main goal of this study was to find out how errors affect user preference of input modality while programming a robot. We were interested in three industrial use cases: assembly, pick&place and welding of points and seams. These use cases were transformed into a simple product manufacturing scenario, better fitting our laboratory settings. A Wizard-of-Oz approach was utilized to avoid implementation specific errors. Without participant's knowledge, a man in a separated room (wizard) observed the scene through a set of cameras and simulated system responses and a feedback. Moreover, WoZ allowed us to simulate certain amount of errors in interaction.

The experimental setup consisted of a table with a top-mounted Kinect v2 sensor and a projector, a robotic platform (PR2) and a touch screen computer besides the table. All sensors were used only for surveillance purposes. During the experiment, the robot was immobile but it helped to create impression of a real robotic workspace.

A simple GUI was created to give users feedback through the projector mounted above the table. There was a bounding box around each object on the table and a label with its name. The selected object was highlighted and points and lines on the objects (selected by a user) were displayed in a different color. The user interface contained a back button used for stepping back, when the system made an error. The button was projected on the table as a red arrow for each modality except the touch screen (there was an on-screen one). Moreover, there was an area dedicated to projecting additional information, animations etc.

A. Input Modalities

Touch table (A) An object is selected by clicking on its projected description. Welding points and seams are selected on a projected image of the object. Assembly constraints are not set with this modality.

Touch screen (B) An object is selected by clicking on it on a screen. Welding points and seams are selected on a zoomed picture of the object. Assembly constraints are not set with this modality. Theoretically there should not be errors in determination of user intention (e.g. where user clicked), but in such a complex system, there could always raise an error, or a user can accidentally click on a wrong place.

Gesture (C) Objects and welding points are selected by pointing on them with the index finger. Welding seams are selected by hovering over a desired seam with the index finger. A gesture used to specify assembly constraint was up to the user. Hand gesture recognition and hand pointing direction recognition is widely studied problem [19], [20]. Recent research shows that 75 to 98% recognition rate is achievable [16], [17].

6D pointing device (D) Similar to C, but instead of the index finger a 6D pointing device was used. Although detection of pose and orientations of this device is more precise and robust than detection of a hand, there still may be errors caused by a user, who can point on a wrong object, or point imprecisely.

Direct robotic arm programming (E) Selecting of objects and welding points and seams was done by pointing on them with a robot's gripper. Just like the 6D pointing device, determining of pose and orientation of a robotic arm is very precise, due to reading arms actuators' internal state, but it can suffer from the same user errors.

Compared to [1], a direct robotic arm programming and a touch table were added. A speech was considered inappropriate as it is probably not sufficiently robust for noisy industrial environments. Our goal was to perform experiment under realistic conditions and we expected participants (mostly university students) to not believe speech programming without predefined vocabulary could work. Moreover, in [1] speech was the lowest rated modality.

Direct robot arm programming (kinesthetic teaching) is commonly used [3], [21], however we are using this technique in a different manner (e.g. selecting objects instead of teaching robot how to grasp them). Touch-sensitive table could be an advantageous alternative to a touchscreen in an industrial environment, as the feedback, system information and interaction with system is held in the user's working space and due to the fact, a user is not forced to divide attention between more places.

B. Tasks

Each participant was told to program the robot to make a simple assembly and packing in a scenario imitating the most common industrial tasks. The scenario was divided into four tasks, each consisting of ten steps (setting ten parameters) in total:

- Assembly: select two objects (e.g. plastic cover and aluminum profile) and set an assembly constraint(s) (e.g. cover orientation)
- Pick&place: select an object and select a place where to put it

- Welding point: select an object, select four points on its top side (to glue stickers in our scenario)
- Welding seam: select an object, select four edges on its top side (to seal boxes with tape)

Each task consisted of ten steps meaning that participant had to set ten parameters: i.e. five times select an object and place where to put it in the pick&place task or select and object and according of its type select one or two assembly constraints in assembly task (see Figure 2). According to participant's group, there were zero, one and three (i.e. 0, 10 and 30%) errors in each task. For instance, in 30% error-level group the system randomly misrecognized three parameters from ten during each of the four tasks. The errors were generated automatically by our WoZ application and were not influenced by the wizard. Order of tasks and steps was the same for all participants.

We see 0% error rate (used for experiment in [1]) as an ideal state however hardly achievable with most of the modalities. 10% seems to be a current realistic level. 30% was selected as the worst case scenario. We assume it to be the worst error ratio probably acceptable by users.

C. Methodology

The SUXES evaluation method for subjective evaluation of multimodal systems has been adopted [22]. It is based on collecting user's expectation and experience and provides means to analyze various interaction methods. The methodology divides experiment into following four phases:

1) Background Information: The experiment is briefly introduced to the subject by a conducter, who is with the subject during the whole experiment. Then, a background information about subject (i.e. age, technical knowledge etc.) is collected.

2) User Expectation: The conductor introduces the shared workspace, all input modalities and the feedback provided by the projector. The subject is allowed to ask questions and to try any modality. Then the subject fills in the questionnaire about his or her expectations based on the introduction.

3) Experiment and User Experience: The conducter guides the subject through four strictly defined tasks: the subject is told what is the current task and step and what to do when error occurs. The task itself is performed solely by the participant. Each subject performs those four tasks with all five modalities (with exception of assembly task, where modalities A and B are skipped). The order of modalities is random for each subject to prevent a learning effect. After that, the subject answers the same questions as in the previous step.

4) Feedback: The subject answers questions about the system using Likert scale rating (see Figures 3 and 4). Most of the subjects also filled valuable fulltext responses.

D. Participants

The experiment has been conducted with 39 participants assigned randomly into three groups. There were eleven males and two females in each group. Participants were mainly university students and researchers with mean age of 23.7 (CI: 22.5 to 24.9) years. Most of them (30) marked themselves as PC experts and at the same time beginners (23) or advanced (15) in robotics. Majority of participants knew what a touchless interface stands for but never used one (31), some indicated that they already used this kind of interface (7) and only one did not know something like this exists.

The whole experiment took approximately 45 minutes for each participant and the interaction itself was recorded by a video camera. Participants' answers have been collected into a spreadsheet.

IV. Results

Participants from all groups (0, 10 and 30% of interaction errors) ordered modalities according to their preference for setting a given parameter before (expectation) and after the experiment (experience). Mean of the order from expectation phase is denoted as r_B and from experience phase as r_A. Statistically significant differences between r_B and r_A within one group were tested using paired t-test (p_{tp}). Differences for a particular modality across the groups were tested using Kruskal–Wallis test with Dunn's multiple comparisons test p_{Wd}. The same test was also used to compare task completion times. Confidence level of 95% was used for all tests. Experience from all participants (all groups) is denoted as r_{Ao}.

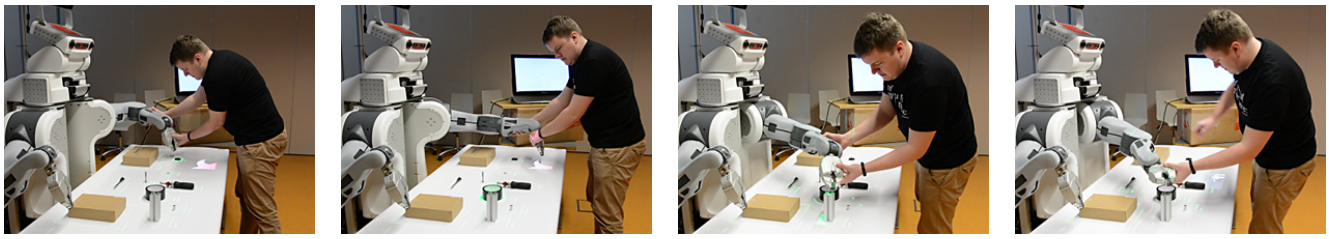
A. Parameters

From the Table I showing users' self-reported data it can be seen for which modality and which parameter there were significant differences between r_B and r_A. Moreover, it can be seen which modality was the most preferred for a given task regardless the amount of errors (r_{Ao}). It should be noted that r_B of C differs between 0% and 30% groups ($p_{Wd} = 0.028$).

Considering the number of significant differences between r_B and r_A from all groups, C and D were ranked significantly better six times, B and E were both worse once and A was worse four times. There are no significant differences in r_B between groups meaning that participants from different groups had similar expectations (with one exception of C in 0% group, parameter *select an object*). Moreover, there are also no differences in r_A. From these results it seems that number of errors in interaction does not have strong impact on preferred modality. In other words, participants from different groups had similar expectations (r_B) as well as experience (r_A). Overall, it seems that participants mostly preferred modalities C, D, followed by A, B and the least preferred was E. Figure 3 shows how participants evaluated expectation and experience for all modalities overall (regardless task).

B. Task Completion Times

Before performing a task the participants were told all relevant information. During the task, only the next step was reminded by the conducter. When beginning the task a participant pressed the "Start" button and then the "Stop" one when finished. We use time between those presses as



(a) User selects plastic cap to be assembled with aluminum profile.

(b) User performs step back as a tape was selected instead of the profile.

(c) Now the intended object (profile) was selected.

(d) Animation shows how the robot understood user's assembly demonstration.

Fig. 2: An example of a typical interaction for the assembly task using the robot arm as an input modality.

modality	group	select an object				select a place				select a point				select a line				assembly constraint			
		r_B	r_A	p_{tp}	r_{Ao}	r_B	r_A	p_{tp}	r_{Ao}	r_B	r_A	p_{tp}	r_{Ao}	r_B	r_A	p_{tp}	r_{Ao}	r_B	r_A	p_{tp}	r_{Ao}
A	0%	3.7	3.3	-	3.3	4.3	3.2	0.015	3.5	3.0	2.9	-	3.2	3.2	3.1	-	3.3	2.8	NA	-	NA
	10%	4.4	3.4	0.012		4.6	3.9	-		3.5	3.1	-		3.8	3.3	-		2.9	NA	-	
	30%	4.1	3.3	0.0024		4.6	3.5	<0.001		2.9	3.5	-		3.0	3.4	-		2.4	NA	-	
B	0%	3.2	2.1	0.02	2.9	2.8	2.1	-	2.7	3.2	2.3	-	2.8	2.9	2.3	-	2.9	3.2	NA	-	NA
	10%	3.7	3.1	-		3.1	2.9	-		3.5	2.9	-		3.7	3.2	-		3.0	NA	-	
	30%	4.2	3.5	-		3.3	3.2	-		3.5	3.2	-		3.2	3.2	-		2.9	NA	-	
C	0%	4.2	4.2	-	3.8	3.7	3.7	-	3.6	2.7	3.7	0.021	3.6	3.0	3.8	-	3.7	3.9	4.0	-	4.1
	10%	2.9	3.6	-		2.9	3.5	-		2.4	3.9	0.0031		2.8	3.9	0.012		3.6	4.5	0.035	
	30%	2.7	3.6	0.046		2.7	3.6	0.027		2.9	3.3	-		3.4	3.4	-		3.8	3.9	-	
D	0%	2.3	3.5	<0.001	3.3	2.5	3.6	0.0045	3.4	3.7	4.0	-	3.7	3.6	3.7	-	3.5	2.3	3.5	0.011	3.4
	10%	2.1	3.4	0.0018		2.5	3.5	0.012		3.6	3.9	-		3.2	3.4	-		1.9	3.9	<0.001	
	30%	2.7	3.1	-		2.9	3.0	-		3.8	3.2	-		3.5	3.5	-		2.2	2.9	-	
E	0%	1.7	1.9	-	1.7	1.7	2.5	-	1.9	2.4	2.1	-	1.7	2.2	2.2	-	1.6	2.8	2.6	-	2.6
	10%	2.0	1.5	-		2.0	1.4	-		2.0	1.3	-		1.6	1.2	-		3.6	2.5	0.021	
	30%	1.4	1.5	-		1.5	1.7	-		1.9	1.9	-		1.9	1.5	-		3.7	2.9	-	

TABLE I: Participants ordered modalities for each parameter separately from the most preferred (5) to the least (1) before (r_B) and after (r_A) the experiment. Where significant difference was found between r_B and r_A p-value is given. r_{Ao} stands for preference after the experiment regardless of the group (0, 10 or 30%).

an objective measure. The Table II shows those times as well as found significant differences between groups for each modality. Differences between modalities are noted below.

The *assembly* task (consisting of *select an object* and *assembly constraint* parameters) was performed only using C, D and E modalities. In all groups there are significant differences between C and E (0%: $p_{Wd} = 0.003$, 10%: $p_{Wd} < 0.001$, 30%: $p_{Wd} < 0.001$) and between D and E (0%: $p_{Wd} = 0.034$, 10%: $p_{Wd} < 0.001$, 30%: $p_{Wd} = 0.002$).

The *pick&place* task consisted of setting *select an object* and *select a place* parameters. In all groups there are significant differences between E and each of rest of the modalities (with max. $p_{Wd} = 0.049$).

The *welding point* task consisted of setting *select an object* and *select a point* parameters. In 0% group, time for B differs from C ($p_{Wd} = 0.0091$) and D ($p_{Wd} = 0.023$). E differs from C and D ($p_{Wd} < 0.001$). In 10% group, time for A, C and D differs from E ($p_{Wd} < 0.001$). The 30% group shows differences between E and A ($p_{Wd} = 0.0018$) and C, D ($p_{Wd} < 0.001$).

The *welding seam* task consisted of setting *select an object* and *select a line* parameters. In 0% group, there is significant difference only between C and E ($p_{Wd} = 0.0029$). 10% group shows difference between E and A, C, D ($p_{Wd} < 0.001$) and 30% group between E and A ($p_{Wd} = 0.0105$), B ($p_{Wd} = 0.014$), C, D ($p_{Wd} < 0.001$).

For most of the tasks C and D were the fastest modalities followed by A and B. E seems to be unsuitable to the sort of tasks as those in this experiment as even 10% of errors affects performance in three of four tasks. It seems that for other modalities a little amount of errors does not play crucial role.

C. System Opinion

The last phase of the SUXES evaluation contains opinion questions. We used the same questions as in [1], with addition of those related to the erroneous behavior (see Figure 4).

Regardless of the group, participants were satisfied with ease of completing the tasks and with time needed to do so. Participants also claimed it was not difficult to understand how to use different modalities. The results are highly similar to those of [1].

mod.	group	assembly		pick&place		welding point		welding seam	
		mean time [s]	significant differences	mean time [s]	significant differences	mean time [s]	significant differences	mean time [s]	significant differences
A	0%	NA	-	34.7 (27.9, 41.5)	0/30: 0.0017 10/30: 0.038	36.8 (31.2, 42.4)	0/30: 0.003 10/30: 0.0022	33.6 (26.1, 41.0)	0/30: <0.001 10/30: 0.0056
	10%	NA		37.4 (33.5, 41.3)		35.2 (31.6, 38.9)		37.0 (33.4, 40.5)	
	30%	NA		47.5 (42.3, 52.6)		49.1 (43.8, 54.4)		53.13 (47.6, 58.6)	
B	0%	NA	-	32.8 (28.7, 36.8)	0/30: <0.001 10/30: 0.04	38.4 (34.6, 42.1)	0/30: <0.001 10/30: 0.0047	36.9 (31.7, 42.1)	0/30: <0.001
	10%	NA		41.2 (36.9, 45.4)		41.6 (37.3, 45.9)		44.2 (41.2, 47.2)	
	30%	NA		52.3 (47.1, 57.4)		54.5 (50.1, 58.9)		53.6 (48.0, 59.3)	
C	0%	54.8 (46.2, 63.5)	0/30: 0.03	28.0 (25.7, 30.3)	0/30: <0.001	28.3 (24.8, 31.8)	0/30: <0.001 10/30: 0.033	27.6 (24.6, 30.6)	0/30: <0.001 10/30: 0.019
	10%	60.4 (47.9, 73.0)		33.7 (29.4, 38.0)		31.9 (27.6, 36.2)		34.8 (30.9, 38.7)	
	30%	70.5 (61.8, 79.2)		40.9 (37.0, 44.8)		41.2 (36.3, 46.1)		47.4 (41.1, 53.7)	
D	0%	61.5 (45.4, 77.6)	0/30: 0.03 10/30: 0.044	28.8 (25.1, 32.5)	0/30: <0.001 10/30: 0.0037	29.3 (25.1, 33.4)	0/30: <0.001 10/30: 0.002	31.0 (26.6, 35.4)	0/30: <0.001 10/30: 0.0023
	10%	61.0 (50.8, 71.2)		32.3 (29.8, 34.9)		31.9 (28.6, 35.3)		36.7 (32.2, 41.1)	
	30%	88.0 (69.3, 106.6)		43.9 (39.1, 48.6)		44.5 (40.6, 48.3)		52.8 (47.9, 57.7)	
E	0%	90.2 (71.2, 109.2)	0/10: 0.013 0/30: <0.001	43.7 (40.5, 46.9)	0/10: 0.0059 0/30: <0.001	42.9 (38.1, 47.7)	0/10: 0.014 0/30: <0.001	42.6 (35.7, 49.4)	0/30: <0.001 10/30: 0.044
	10%	129.6 (112.2, 146.9)		60.6 (54.9, 66.2)		58.9 (54.1, 63.6)		58.4 (52.7, 64.1)	
	30%	156.7 (127.6, 185.8)		75.3 (69.2, 81.4)		83.0 (68.2, 97.8)		85.1 (68.9, 101.3)	

TABLE II: Task completion mean times (with 95 % confidence intervals) for all modalities, groups and tasks. For each modality, significant differences between times are noted where found in form of $group_x/group_y : p_{Wd}$.

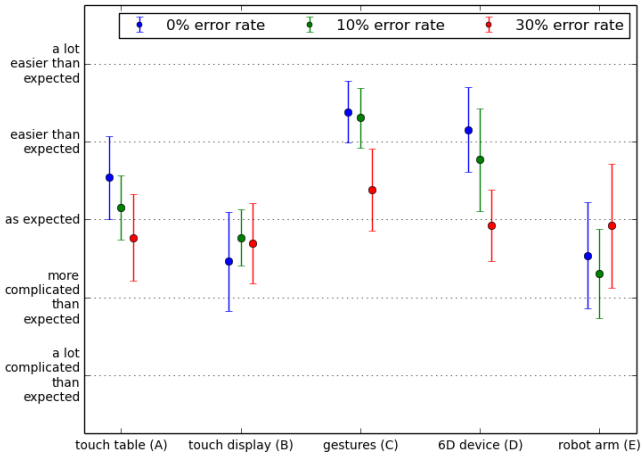


Fig. 3: User's assessment how experience matched expectation.

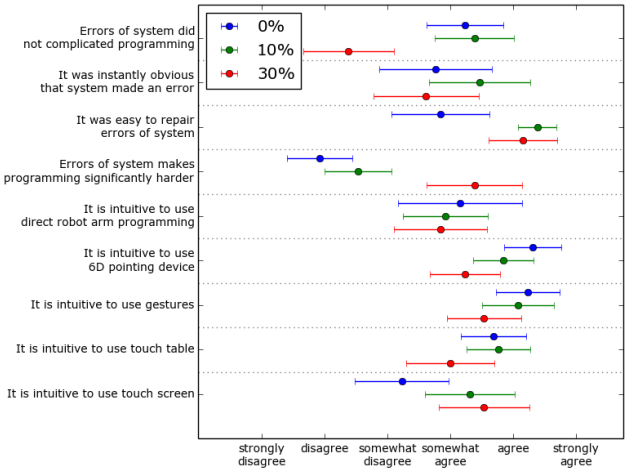


Fig. 4: System opinion

Most of the subjects rated modalities C and D similar, however had a stronger believe in 6D pointing device as they expect it to be more precise than gesture, despite there was the same amount of errors. Participants were also often distracted by the fact, that feedback was always projected on the real objects on the table and not on the place they were working with. Especially, for B most of them would prefer feedback (e.g. selected object) to be shown on the screen and not only on the table. This was however done by purpose, to ensure each modality has exactly the same feedback and participants were noticed about this in advance.

In questions related to erroneous behavior a difference can be seen between error groups. With a growing amount of the errors, perceived intuitiveness of the modalities decreases, except for the touch screen, where it grows (see Figure 4). This could be caused by the fact, that the touch screen is the only control commonly used by the participants. Moreover, the back button was on the screen, so the participants were not forced to think about how to press projected button as for other modalities. Modalities B and E were in general evaluated as the least intuitive. Participants stated that with growing amount of errors, programming was significantly harder and that errors in communication complicated programming.

A few of the participants found out that errors were made by purpose or that some parts of system were simulated. However, according to feedback and discussion with participants, none of them found out the experiment was WoZ.

V. Conclusions

The aim of the conducted experiment was to explore how different modalities used for setting common parameters

when programming a robot cope with interaction errors. Participants were divided into three groups according to amount of simulated errors. Their ranking of the modalities before and after the experiment as well as answers from feedback phase were analyzed as subjective measures. Moreover, task completion times were recorded and analyzed as an objective measure.

The gesture and 6D pointing device modalities were the most preferred and fastest modalities in all groups. Touch-sensitive table and display were in general preferred similarly and similar task completion times were obtained. With respect to the task completion times as well as feedback from participants (system opinion) the robot arm seems to be inappropriate as a pointing device for tasks as those in this study and its usage should be reconsidered. It seems that order of preferred input modalities for a given task is not affected by amount of interaction errors. Obtained results support our prior speculation of 10% to be an acceptable level of errors and 30% to be a worst case scenario as especially task completion times grow dramatically.

According to the results, multi-modal interaction based on gestures with complementary usage of a 6D pointing device seems to be promising. We also see touch-sensitive table as a perspective modality however it will be necessary to improve interaction and solve setting more complicated parameters as the *assembly constraint*. The robot arm has advantage of no additional cost however, its usage is physically more demanding than other modalities and for our use-case with relatively simple tasks it had no added value. However, for different types of tasks, e.g. requiring high precision, it could be more useful.

It should be noted that our study simulated the same amount of errors for all modalities. In practice, it can be expected that for instance robot arm modality will be less error-prone than gesture recognition.

As a future work, we will extend the ARTable prototype. The projected interface will provide more information and be fully interactive in conjunction with a touch-sensitive table. Instead of a touch display, a hand-held device or a see-through video glasses with augmented reality will be used. We will also experiment further with robot arm as it could be useful for complex tasks.

ACKNOWLEDGMENTS

This work was supported by The Ministry of Education, Youth and Sports of the Czech Republic from the National Programme of Sustainability (NPU II); project IT4Innovations excellence in science - LQ1602.

REFERENCES

[1] S. Profanter, A. Perzylo, N. Somani, M. Rickert, and A. Knoll, "Analysis and semantic modeling of modality preferences in industrial human-robot interaction," in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ Int. Conference on*, Sept 2015, pp. 1812–1818.

[2] A. Billard, S. Calinon, R. Dillmann, and S. Schaal, "Robot programming by demonstration," in *Springer Handbook of Robotics*. Springer, 2008, pp. 1371–1394.

[3] C. Schou, J. S. Damgaard, S. Bogh, and O. Madsen, "Human-robot interface for instructing industrial tasks using kinesthetic teaching," in *Robotics (ISR), 2013 44th Int. Symposium on*. IEEE, 2013, pp. 1–6.

[4] S. Alexandrova, Z. Tatlock, and M. Cakmak, "Roboflow: A flow-based visual programming language for mobile manipulation tasks," in *Robotics and Automation (ICRA), 2015 IEEE Int. Conference on*. IEEE, 2015, pp. 5537–5544.

[5] F. J. Abu-Dakka, B. Nemec, A. Kramberger, A. G. Buch, N. Krüger, and A. Ude, "Solving peg-in-hole tasks by human demonstration and exception strategies," *Industrial Robot: An Int. Journal*, vol. 41, no. 6, pp. 575–584, 2014.

[6] A. Perzylo, N. Somani, S. Profanter, M. Rickert, and A. Knoll, "Multimodal binding of parameters for task-based robot programming based on semantic descriptions of modalities and parameter types," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS), Workshop on Multimodal Semantics for Robotic Systems, Hamburg, Germany*, 2015.

[7] J. Lambrecht, M. Kleinsorge, M. Rosenstrauch, and J. Krüger, "Spatial programming for industrial robots through task demonstration," *Int J Adv Robotic Sy*, vol. 10, no. 254, 2013.

[8] J. Norberto Pires, J. Norberto Pires, G. Veiga, and R. Araújo, "Programming-by-demonstration in the coworker scenario for smes," *Industrial Robot: An Int. J*, vol. 36, no. 1, pp. 73–83, 2009.

[9] K. R. Guerin, S. D. Riedel, J. Bohren, and G. D. Hager, "Adjutant: A framework for flexible human-machine collaborative systems," in *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ Int. Conference on*. IEEE, 2014, pp. 1392–1399.

[10] A. Gaschler, M. Springer, M. Rickert, and A. Knoll, "Intuitive robot tasks with augmented reality and virtual obstacles," in *Robotics and Automation (ICRA), 2014 IEEE Int. Conference on*. IEEE, 2014, pp. 6026–6031.

[11] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from rgb-d videos," *The Int. J of Robotics Research*, vol. 32, no. 8, pp. 951–970, 2013.

[12] G. Hirst, S. McRoy, P. Heeman, P. Edmonds, and D. Horton, "Repairing conversational misunderstandings and non-understandings," *Speech communication*, vol. 15, no. 3, pp. 213–229, 1994.

[13] A. B. Beck, A. D. Schwartz, A. R. Fugl, M. Naumann, and B. Kahl, "Skill-based exception handling and error recovery for collaborative industrial robots," in *Procs. FinE-R Workshop*, 2015, pp. 5–10.

[14] C. Breazeal, C. D. Kidd, A. L. Thomaz, G. Hoffman, and M. Berlin, "Effects of nonverbal communication on efficiency and robustness in human-robot teamwork," in *Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ Int. Conference on*. IEEE, 2005, pp. 708–713.

[15] S. R. Klemmer, A. K. Sinha, J. Chen, J. A. Landay, N. Aboobaker, and A. Wang, "Suede: a wizard of oz prototyping tool for speech user interfaces," in *Procs. of the 13th annual ACM symposium on User interface software and technology*. ACM, 2000, pp. 1–10.

[16] M. Pateraki, H. Baltzakis, and P. Trahanias, "Visual estimation of pointed targets for robot guidance via fusion of face pose and hand orientation," *Computer Vision and Image Understanding*, vol. 120, pp. 1–13, 2014.

[17] D. Shukla, O. Erkent, and J. Piater, "Probabilistic detection of pointing directions for human-robot interaction," in *Digital Image Computing: Techniques and Applications (DICTA), 2015 Int. Conference on*. IEEE, 2015, pp. 1–8.

[18] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[19] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: a survey," *Artificial Intelligence Review*, vol. 43, no. 1, pp. 1–54, 2012. [Online]. Available: http://dx.doi.org/10.1007/s10462-012-9356-9

[20] J. Suarez and R. R. Murphy, "Hand gesture recognition with depth images: A review," in *RO-MAN, 2012 IEEE*, Sept 2012, pp. 411–417.

[21] S. Alexandrova, M. Cakmak, K. Hsiao, and L. Takayama, "Robot programming by demonstration with interactive action visualizations," in *Procs. of Robotics: Science and Systems*, Berkeley, USA, July 2014.

[22] M. Turunen, J. Hakulinen, A. Melto, T. Heimonen, T. Laivo, and J. Hella, "Suxes-user experience evaluation method for spoken and multimodal interaction." in *INTERSPEECH*, 2009, pp. 2567–2570.