Int J Soc Robot (2013) 5:477–490
DOI 10.1007/s12369-013-0204-0

Assessing Interaction Dynamics in the Context of Robot Programming by Demonstration

Ana Lucia Pais · Brenna D. Argall · Aude G. Billard

Accepted: 22 June 2013 / Published online: 2 August 2013

Abstract In this paper we focus on human–robot interaction peculiarities that occur during programming by demonstration. Understanding what makes the interaction rewarding and keeps the user engaged helps optimize the robot's learning. Two user studies are presented. The first one validates facially displayed expressions on the iCub robot. The best recognized displays are then used in a second study, along with other ways of providing feedback during teaching a manipulation task to a robot. We determine the preferred and more effective way of providing feedback in relation to the robot's tactile sensing, in order to improve the teaching interaction and to keep the users engaged throughout the interaction.

Keywords Robot programming by demonstration · Robot facial displays · Emotion expression · Interaction dynamics · Incremental learning

1 Introduction

Programming by Demonstration (PbD) methods contribute to Human–Robot Interaction (HRI), by making robots accessible to naive users, who have little knowledge of a robotic platform or programming language. Necessary tools are provided so that a robot is able to learn how to accomplish a task by simply observing the necessary gestures. This paper focuses on evaluating the user-friendliness of our framework for teaching a robot how to refine its manipulation skills [26]. Specifically we seek to identify the factors that make the interaction more engaging for the teacher. An engaged user may be more willing to teach the robot longer, and may pay more attention to the procedure, which may improve the robot's performance [11].

Evaluating a robot teaching by demonstration procedure can be done with respect to

(1) *the quality of the demonstration* as a measure of the amount of useful data that can be included in learning the task [26];
(2) *the teaching efficiency*, which is a measure of how well the robot can reproduce the demonstrated task [7] and
(3) the *perceived user satisfaction*, which is the aspect addressed in this paper.

The framework that we are evaluating consists of a multi-step iterative learning procedure, in which a human shows a robot multiple ways of holding a can, via tactile feedback, and several rounds of demonstration. The teaching procedure consists of three phases:

(1) *Demonstration*. The user shows the robot different ways of holding an object by moving the robot's fingers, using their passive compliance capability. A certain contact signature corresponds to each demonstrated posture, and is reflected by the activation of the robot's tactile sensors on the fingertips.
(2) *Replay*. The robot replays the sequence of hand postures, to record data that is not influenced by the touch of the teacher. Finally, we have

A.L. Pais (✉) · A.G. Billard
Learning Algorithms and Systems Laboratory, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
e-mail: lucia.pais@epfl.ch

A.G. Billard
e-mail: aude.billard@epfl.ch

B.D. Argall
Departments of EECS and PMR, Northwestern University, Chicago, IL USA
e-mail: brenna.argall@eecs.northwestern.edu

(3) *Testing*. The adequacy of the learned model is reflected by the robot's ability to adapt the fingers' positions in response to perturbations in the position of the object. Alongside the teaching procedure users are provided with various feedback modalities (detailed in Sect. 4) that expresses the robot's current state. The following section reviews works on identifying human factors involved in HRI teaching applications, which are the basis of the work presented here. Section 3 presents user study results validating a set of facial expressions on the humanoid robot iCub, which are later used as feedback in our framework. Section 4 describes our PbD interface and assesses the HRI development during teaching. Section 5 presents conclusions.

2 Related Work

From a human perspective, teaching a robot by demonstrating a task is a natural approach as it resembles the way humans teach another person [16, 20]. From a robot's perspective, learning can occur (a) by observing gestures, natural language, and other cues offered by the teacher or (b) by experience, being directly guided through the task.

Natural methods for robot task learning include [19]: instructive demonstrations, generalization over multiple demonstrations and practice trials. In our work we take a similar approach by including demonstrations, rounds of replay, and testing. These guidelines are complemented by stressing the importance of using social cues as a natural way of structuring and guiding the robot's learning [4]. The robot should make its states transparent to the tutor by using communicative acts, while the instructor builds a mental model of what the robot has learned. While this highlights the importance of bi-directional teaching [9], which allows for the improvement of both learner and teacher, it also raises two main concerns:

(1) finding the appropriate type of feedback for the robot to provide so that the teacher easily understands the effects that teaching has on the robot and
(2) designing the interaction so that the tutor does not lose interest in teaching.

To address the first question, various ways of providing feedback in tutoring applications have been tested: *gazing* at what the teacher is doing [4]; *emotional reactions* that influence human performance in collaborative tasks [27]; *verbal cues* that increased the frequency and accuracy of demonstrations in a dancing task [17]. Given that proper feedback is provided, the social component goes as far as attributing emotional states to artificial objects [12], thus increasing the user's implication. In our case, holding an object requires good contact on all fingertips and, in particular, on fingers placed in opposition on the object to ensure the stability of the grasp. Therefore, we take a similar approach to [4] and make this information (i.e. how good the contact is at the fingertips) transparent to the user, by correlating it with different feedback modalities. This helps the user create a mental model of the level of adaptation the robot achieves throughout multiple rounds of demonstration, replay and testing.

Addressing the second question of whether the interaction is sustainable is particularly relevant in demonstrating a task to a robot because the user should be engaged for the proper amount of time to deliver the required number of demonstrations. The initial interaction might be driven by curiosity [13], and perceiving the robot as "intelligent" due to its vividness might keep the user interested in the interaction [2]. But a sustained interaction is subject to six factors [22] responsible for keeping the user engaged. The first two factors, described in [22], address the problem of setting up the interaction, by: (1) providing contextual objects and knowledge, shown to dramatically improve human participation, as well as (2) initiating the interaction.

The other four factors focus on regulating the interaction by:

(3) having the robot provide responses in a timely manner and having a mechanism for managing role-switching,
(4) using feedback to express robot's states,
(5) using turn taking for sustaining a certain rhythm in the interaction and
(6) confirming robot's engagement by showing attention.

These factors increase the complexity of the interaction, which may promote accepting the robot as an interaction partner [10].

In our work we aim to add social components to a programming by demonstration interaction such that it keeps the user engaged and willing to deliver better quality demonstrations, see Experiment II. In designing the interaction we use four out of the six factors mentioned above, throughout the whole teaching procedure: first, the user is given contextual knowledge about the task to be performed; second, the robot responds in a timely manner to the user's actions; third, the teaching procedure is implicitly designed for turn taking by alternating the user's lead in the demonstration and testing phases with the robot's lead in the replay step; and fourth, the robot's states are conveyed to the user.

We test three active ways in which to convey the robot's internal states, namely via verbal feedback from a knowledgeable person, a graphical user interface and robot facial expressions. These modalities are contrasted against a control group in which no feedback was offered. For using adequate expressions a prior user study is conducted to validate a set of 20 custom face displays and choose the best

recognized ones, see Experiment I. Adding social components to the teaching paradigm [4, 6, 7], changes the classical approach to teach robots, where the robot is passive and learns solely from observing the teacher performing the task. The active feedback provided by the robot contributes to human–robot team work, having both agents work cooperatively to achieve the same goal, namely transfer of skills.

3 Experiment I

Validating a robot's expressive capabilities is a necessary step before using them in real applications, as embodiment particularities can influence both the way the user perceives the expressions and the recognition accuracy [1]. Thus we conducted an experiment to assess to what extent humans can decode and interpret facial emotion expressions on the iCub robot. The goal was to determine a subset of best recognized expressions that we could later use to provide feedback in a PbD framework, described in Sect. 4. The underlying model for building the emotional displays and the implementation are described next.

3.1 iCub Facial Displays

Emotion Representation When using robot emotions it is important to represent them in a way humans could easily understand. Russell [23] determined that humans have an innate capability of representing affect and thus proposed a circumplex model of clustering emotions, containing 28 facial expressions positioned in a two dimensional space. The first dimension emerges in studies of intra-personal behavior, and it is easily interpretable regardless of the users' culture, while the second dimension is validated on interpersonal behavior [24]. The dimensions are considered implicit in the human understanding of emotion [24] and are given by (a) valence, pleasure or positivity and (b) activity, arousal or activation [24]. Our work will refer to this first axis as *valence* and the second as *arousal*.

The design of emotion displays used in this study was based on Russell's model of arousal and valence [25] because:

(1) it provides an easy mapping between emotion features and robot expressive capabilities,
(2) these dimensions are easily interpretable as discussed above, and
(3) these dimensions emerge in inter-personal behavior, making the emotions validated in this study suitable for communicating internal states in HRI. In robotics applications, the arousal and valence dimensions are explored in different contexts. The first dimension can be communicated through haptic interaction [28], while the emotional valence of a situation can lead to perceiving a robot as being empathetic [8].

Expressions Implementation The facial expressions were implemented on the humanoid robot iCub using LEDs for representing the eyebrows and mouth, and actuators for controlling the eyelids opening angle. The changes along the arousal dimension were modeled by the opening of the eyelids and the curvature of the eyebrows, while the changes along the valence axis were mapped to changes in the lip curvature. LEDs are used to project the eyebrows and mouth facial features onto the face shell. The projection makes the line of consecutive individual LEDs appear continuous. There are 19 LEDs for the mouth and four sets of five LEDs for the eyebrows.

3.2 Study Design

A subset of 20 out of 28 expressions in Russell's original model were chosen arbitrarily as representing the maximum set of iCub displays that could be easily distinguishable. The designed expressions fit two valence levels (positive and negative) and three arousal levels (low, medium, and high). The displays were investigated, according to four categories:

(1) positive, and intense: astonishment, delightedness, gladness, happiness, and pleased;
(2) negative, and intense: alarmed, afraid, tensed, angry, and annoyed;
(3) negative, not intense: miserable, depressed, sad, gloomy, and bored;
(4) positive, not intense: satisfied, content, serene, calm, relaxed.

This way of dividing emotions allowed us to assess the degree of granularity that we could use for the expressions to still be interpretable by the users. Thus we evaluated the recognition rates on different levels of granularity: two classes, if only the distinction between positively and negatively experienced emotions was considered, three classes according to the arousal levels; four classes, given by Russell's categories and 20 classes when classification by emotion name was considered.

The study addressed the overall question of how easily the iCub's facial expressions could be recognized if conveyed only through features like lip curvature, eyebrows and eyelids. We made the following untested assumptions:

(1) the designed mapping between human emotions and robot displays was correct, implying that the implemented expressions were as close as possible to the human ones;
(2) subjects were able to identify these emotions in humans.

Based on these assumptions, our working hypotheses were:

H1: The categories in Russell's model of emotions are identifiable in robot expressions by most humans.
H2: Subjects claiming to be skilled in recognizing human emotions might also be skilled in recognizing robot displays.
H3: The time a user requires for classifying an emotion is correlated with the arousal level of that emotion.

Participants The experiment involved 23 participants (five females and 18 males), from various places of origin (13 European, six Asian, four North American), with an average age of $M = 27.52$, standard deviation $SD = 5.43$ (minimum of 21 and maximum of 48).

Study Protocol In a pre-experiment questionnaire the subjects had to assess their skill in understanding human emotions. The questions were:

1. How often can you read a person's facial expressions? (Never/rarely/often/always)
2. How often do you check for emotional cues while interacting with a person? (Never/rarely/often/always)
3. What is easier for you to recognize from a person's facial expression? (Sadness/happiness/both)

The answer to each of the first two questions was marked with a score from 0 to 3, for the third question a point was given for being able to recognize sadness or happiness, two points for both or minus two for none. The sum of the points obtained represented a general evaluation of the responders' confidence levels (self-assessed skill) in recognizing human emotions. Based on this score participants were divided in three skill levels: low, four subjects; medium, nine subjects; and high, ten subjects.

In the second part of the study, the subjects were shown the facial displays, and for each asked to: classify the display as positive or negative valence, to assign an arousal level, and a name from a given list, and to rate the arousal level in comparison to the previous emotion. Each participant was exposed to a sequence of 60 facial displays, consisting of 20 different expressions, each repeated three times. The order in which the expressions were displayed was randomized, while avoiding the consecutive display of identical or closely related emotions. Participants were facing the robot during the whole experiment. The subject controlled the moment when the displayed emotion changed. They were not shown examples of iCub facial expressions prior to taking the survey. The time between the emotion display and the selection of each answer was recorded. Participants were not told that the experiment was timed, to avoid rushed answers. The survey required up to 40 minutes per user for completion. The study language was English, however, as not all subjects were native speakers, some required clarifications for emotion names. Commonly hard to distinguish emotion terms were "content vs. serene"; "calm vs. relaxed"; and "sad vs. gloomy".

In a post-experiment questionnaire the subjects were asked to rate their general expectations of HRI when these facial displays would be provided. On a five level Likert scale [18] subjects rated the *Interaction* (ranging from distracting to engaging), and the *Aesthetical* component (ranging from unpleasant to pleasant).

Measurements The coding of each emotion was done using an initially assigned value for valence (P = positive or N = negative), one of three arousal levels (L = low, M = medium, and H = high), and a name label, based on Russell's mapping of emotions to the arousal and valence axes (see Table 1, columns 1 and 2). For each facial emotional expression we recorded the arousal, valence levels and the name label attributed by the user, and the time the user took to assign a value. Secondly we recorded the user's answers to the pre and post-experiment questionnaires.

3.3 Results

Results are presented in relation to the working hypotheses, and consist of evaluating the recognition rates for each emotion, and in making a subjective evaluation of the user's experience while seeing the displays.

3.3.1 Recognition Rates

To determine whether participants were able to correctly identify emotions in Russell's model (hypothesis H1), recognition rates were evaluated over multiple categories, in order to assess how well people can differentiate between different levels of granularity. Recognition rates for the valence level (Table 1, third column), arousal level (fourth column) and name (sixth column) were computed by comparing the score attributed by the user for each level and the name label with the initially assigned values for each emotion. A good match was marked with 1 and a no-match with 0. The rates presented in Table 1 represent the percentage of recognized displays (number of matches) from the 60 total displayed emotions. Similarly, the recognition score for both valence and arousal levels (fifth column), represents the number of correct matches for both levels, from the total number of displayed emotions.

Recognition rates vary across categories (see Fig. 1(a)). The best recognized emotion from each category is shown in Fig. 2. Recognition rates for positive emotions tended to decrease as the arousal increased, while with negative emotions, the opposite trend was observed. Participants could identify the emotion valence (positive vs. negative)

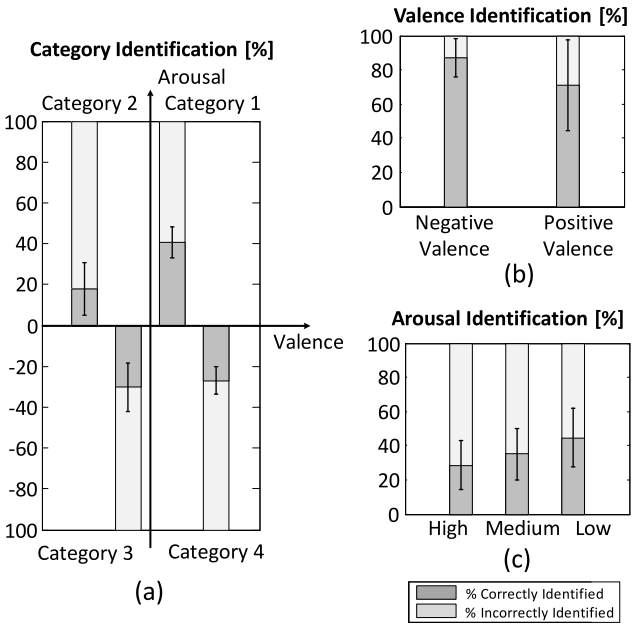

Fig. 1 Recognition rates [%] for: (**a**) Russell's categories, (**b**) valence, (**c**) arousal levels

Table 1 Percentage of correctly identified emotions by valence and arousal levels, by both arousal and valence, and by name. The coding indicates a positive (P) or negative (N) valence and low (L), medium (M) or high (H) arousal level

Coding		Emotion	Valence [%]	Arousal [%]	Both [%]	Name [%]
1.	P_H	Astonish	68.12	44.93	31.88	39.13
2.	P_H	Delight	89.86	36.23	36.23	11.59
3.	P_M	Glad	89.86	50.72	50.72	07.25
4.	P_M	Happy	91.3	47.83	46.38	15.94
5.	P_M	Pleased	86.96	42.03	39.13	10.14
6.	N_H	Alarmed	63.77	13.04	08.70	11.59
7.	N_H	Afraid	92.75	20.29	20.29	0
8.	N_M	Tense	81.16	43.48	39.13	02.90
9.	N_M	Angry	85.51	13.04	13.04	76.81
10.	N_M	Annoyed	98.55	10.14	08.70	10.14
11.	N_M	Miserable	95.65	23.19	21.74	15.94
12.	N_M	Sad	89.86	47.83	44.93	17.39
13.	N_L	Gloomy	88.41	37.68	30.43	05.80
14.	N_L	Bored	73.91	56.52	40.58	18.84
15.	N_L	Depressed	98.55	13.04	13.04	14.49
16.	P_M	Satisfied	84.06	43.48	36.23	05.80
17.	P_M	Content	95.65	30.43	30.43	10.14
18.	P_L	Serene	34.78	50.72	24.64	02.09
19.	P_L	Calm	39.13	49.28	18.84	15.94
20.	P_L	Relaxed	28.99	60.87	24.64	07.25

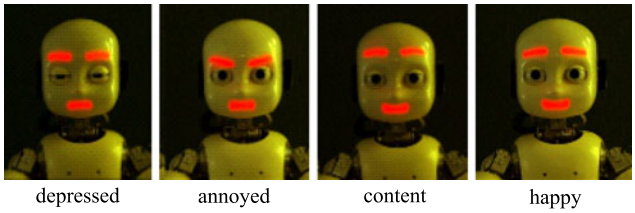

Fig. 2 The best recognized facial displays with respect to the valence level from each of the four categories

for more than two thirds of the emotions ($M = 78.84$ %, $SD = 21.34$ %); see Fig. 1(b). This correlates well with the fact that all participants agreed that they were capable to recognize when someone was happy. Similarly, participants correctly associated Depressed, Miserable and Sad with a negative emotion, even though they did not always label the displayed emotion correctly. This again correlates well with participants' ability to recognize when someone was sad. Analysis of recognition rates for each of the three arousal levels (see Fig. 1(c)) shows that participants had a tendency to better recognize low arousal ($M = 43.47$ %, $SD = 19.14$ %) and medium arousal ($M = 35.93$ %, $SD = 15.63$ %) emotions, than high arousal emotions ($M = 28.62$ %, $SD = 14.56$ %). In other words, the less intense the emotion (whether positive or negative), the better it was recognized. This observation did not seem to match the observation that participants were good at recognizing positive vs. negative emotions, and generally at associating emotions to the correct Russell category. We suspect that these poor results are due to the fact that participants may confuse some closely related emotions. The confusion matrix for the intensity levels showed that in 53.62 % of the cases the negative-medium emotions were mistaken for negative-high emotions, and positive-low for negative-low (18.55 %), while negative-low emotions were equally assigned to negative-low or negative-medium. The name recognition rates for each emotion showed rather poor results, with an average of 20 %. This is partially justified by the difficulties subjects had in understanding the different terms used for the given emotions.

Results presented in this subsection partially support hypothesis H1 for low levels of granularity (e.g. differentiating positive emotions vs. negative displays). While category recognition rates were above chance level (5 %), they were overall poor. This is likely due to the simplicity of the LED coding, which does not allow rendering the full complexity of human facial expressions.

3.3.2 Human Factors Influence on Recognition Rates

To test hypotheses H2 and H3, we tested the influence of human factors on the recognition rates, mainly the user's self-assessed *skill* in recognizing human emotions, the *reaction*

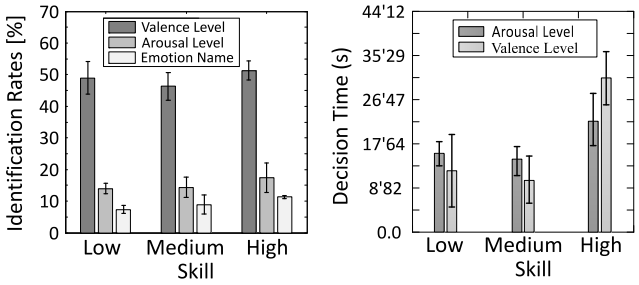

Fig. 3 Effect of the user self-assessed skill of recognizing emotions on (*left*) the emotion arousal level and (*right*) the time necessary to assign arousal and valence levels to a facial display

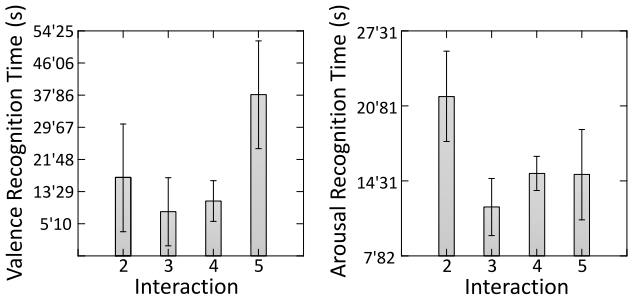

Fig. 4 Effect of the perceived interaction (ranging from 1 (distracting) to 5 (engaging) on the recognition time for (*left*) emotion valence (*right*) emotion arousal level

times (the time necessary to assign the appropriate levels to each displayed emotion), and the user-perceived *aesthetics* of the displays.

A. Evaluation of User's Skill We hypothesized that if participants felt confident in their general ability to assess emotions, they would also be more competent at recognizing robot emotions. Thus, we made a more general assessment about how confident participants were at recognizing emotions in general. Almost half of the participants declared themselves as confident in their ability to detect a sad person ($M = 52.17$ %, $SD = 0.51$ %). The vast majority of participants claimed to be able to recognize when a person was happy ($M = 82.60$ %, $SD = 0.38$ %). Most participants declared that they were often able to recognize facial expressions and they often searched for facial cues while interacting with a human partner ($M = 82.60$ %, $SD = 13.27$ %).

We tested the influence on the category-based recognition rates of three factors[1] that aimed at significant effects:

(1) skill ($F(2, 1379) = 69.9$, $p = 0.001$),

(2) valence level ($F(1, 1379) = 4.15$, $p = 0.04$) and

(3) arousal level ($F(2, 1379) = 3.04$, $p = 0.01$).

The recognition rates are presented in relation to the three levels of skill in Fig. 3(a). The users' self-assessed skill in recognizing human emotions was not correlated with the recognition rates, showing that hypothesis H2 was not supported.

The degree of engagement that the users assign to the human–robot interaction when facial cues are involved is correlated with the recognition rates. Thus, people who rated the robot-expressed emotions as being very engaging were also good at recognizing emotions ($F(3, 1316) = 98.124$, $p < 0.01$). The effect of how aesthetic the interaction is when facial expressions are used is also significant ($F(4, 1315) = 50.96$, $p = 0.001$). Age was also found to have a significant impact on identifying the emotion valence, ($F(1, 1369) = 98.575$, $p = 0.001$), and arousal level ($F(2, 1369) = 164.784$, $p = 0.002$), showing that identification rates decrease with age.

B. Evaluation of Users' Reaction Times We tested the effect of three factors on users' reaction times: the emotions' arousal and valence levels and users' skill. The average time required to classify valence was 10.41 s for negative emotions and 16.7 s for positive emotions, suggesting that negative emotions were easier to understand. The average time necessary for assigning an arousal level was significantly lower for high arousal emotions (10 s) compared to low arousal emotions (20 s). The arousal level had a significant impact on the time the user took to rate the displayed emotion ($F(2, 1375) = 10.34$, and $p = 0.002$). Skill, however, did not have a significant effect on the arousal level classification time, but only on the valence classification time ($F(2, 1377) = 5.495$, $p = 0.004$); see Fig. 3(b). Average valence identification time for people that consider themselves not skilled in recognizing human emotions was 10 s, while for high skilled people was almost 30 s, suggesting that people who considered themselves skilled in recognizing human emotions might be more motivated during the interaction. In addition, users that rated the interaction as engaging took a longer time to recognize if an emotion was positive or negative (Fig. 4(a)), but had better recognition times for emotion arousal level than those who rated the interaction as distracting (see Fig. 4(b)).

Hypothesis H3, stating that the time to decision required for classifying an emotion into a category was negatively correlated with the arousal level of that emotion, was supported by the results presented in this subsection.

C. User-perceived Aesthetic Component In the last part of the experiment, participants were asked to rate the *aesthetics* of the interaction (ranging from unpleasant to pleasant) when robot facial displays were provided. The aesthetics component was rated lowest by persons that rarely check for expressions of emotion in humans (2 subjects). The highest

[1] Analysis was based on ANOVA, a statistical technique used for testing the null hypothesis that there is no difference between groups. It is based on comparing the mean value of a common component. When the null hypothesis is false, the result is significant, implying an F value greater than 1, and a p-value $p \leq \alpha$, e.g. $\alpha = 0.05$.

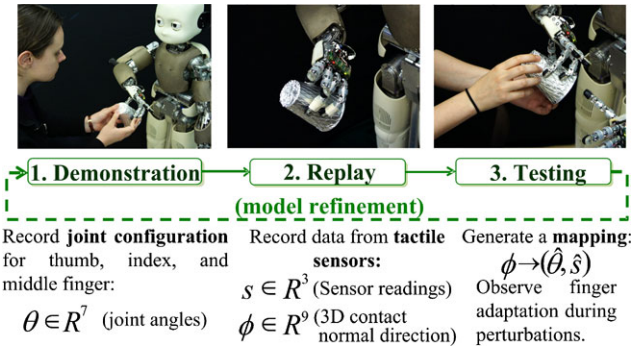

Fig. 5 PbD framework for teaching a manipulation task to a robot. A human shows the robot various ways of holding a can through tactile guidance. The robot replays the demonstrated motion and learns a model of the task that can be tested and further refined by providing additional demonstrations

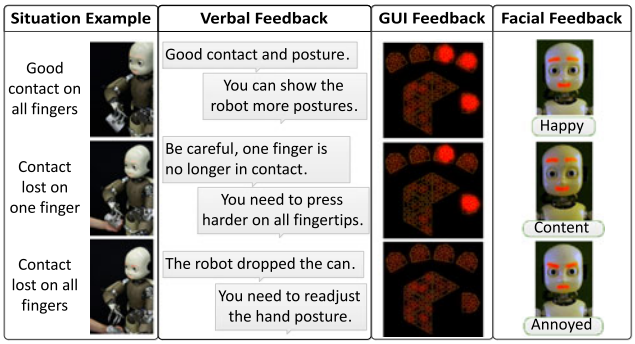

Fig. 6 The usability of the PbD framework was tested using different modalities of providing feedback to the user. The experimental setups according to the three major situations in the experiment: verbal feedback, GUI feedback, facial feedback. These setups were contrasted against a no-feedback situation

rating was given by the group of subjects that always check for emotional expressions in other persons (16 participants). This group also had the best recognition rates for valence ($M = 53$ %, $SD = 0.2$) and arousal levels ($M = 19.6$ %, $M = 0.8$).

Overall, the above chance recognition rates occurred for all categories, with the best rates found for the smallest level of granularity (i.e. classification in two classes, positive and negative emotions).

4 Experiment II

The second experiment was carried out to study the impact of providing different types of robot feedback on the effectiveness of a teaching by demonstration framework, as well as on user satisfaction. The goal is to have human-users, with no prior experience of interacting with the iCub platform, be able to teach the robot how to refine its manipulation capabilities and achieve a satisfactory model of holding a certain object, after providing the robot with several rounds of kinesthetic teaching.

A multi-step training procedure, illustrated in Fig. 5 was used to iteratively build a training data-set from teacher's demonstrations and learner's replay. The teaching procedure consisted of three steps. The first step was the *demonstration*, in which the user demonstrated the robot different finger positioning on the object using tactile guidance. The robot held the object with three fingers of the right arm (the thumb, index and middle finger), maintaining contact just on the fingertips. The information recorded in this step consisted of a set $\Theta \in R^7$ of robot finger joint angles. The second step was the *replay*, in which the robot replayed the demonstrated motion in order to record for each posture the corresponding tactile-sensor signature, without being influenced by the additional pressure provided by the teacher. The contact information was recorded using the pressure response of the tactile sensors on the robot's fingertips. Each fingertip has 12 tactile nodes that were activated on contact with the object, providing an 8-bit pressure value. Information recorded at this stage consisted of sensor readings $s \in R^3$, representing an averaged value for each fingertip, and a vector $\phi \in R^9$ representing the computed 3D contact normal direction. Based on the information recorded in the first two steps, the robot used statistical techniques to learn a mapping between the tactile response on its fingers and the corresponding finger positions $\phi \rightarrow (\hat{\Theta}, \hat{s})$, as described in [26]. When a perturbation occurred the contact signature changed. The learned model allowed the robot to predict a new hand configuration based on the new sensed contact. The third step was the *testing*, in which the participant could test the learned model by perturbing the position of the object. The displacement of the robot's fingers in response to perturbation gave an indication of the adequacy of the model. The obtained model could be further refined by providing additional rounds of demonstration, replay and testing. The tactile information was important because of the way it was accounted for in the learning algorithm. According to the reliability measure introduced in [26] the stronger a contact sensor reading was, the more reliable it was considered to be. This implied discarding weak contact readings. Thus providing the user with a valid representation of this information would dramatically improve the amount of useful information provided through demonstration. This would be reflected in the learned model by achieving better adaptation.

4.1 Study Design

The experiment was performed on the iCub robot. We studied four conditions (experimental setups), shown in Fig. 6, which reflected the type of feedback being provided. In the first setup (**E1**) no feedback was provided by the robot, nor

by the experimenter. This setup was called *no feedback*. In the second setup (**E2**) rich verbal feedback was given by a knowledgeable experimenter whenever it was considered necessary (*verbal feedback*). In the third setup (**E3**), a Graphical User Interface (GUI) was used consisting of a diagram of the tactile nodes on each fingertip. The GUI provided a real-time continuous feedback on the tactile sensing intensity and area of activation, by highlighting the activated tactile nodes. The subject knew when the object was in contact with the robot's fingertips and could see the variation in the contact area (*GUI feedback*). In the last setup (**E4**), robot facial expressions were provided as discretized feedback to the subject on the adequacy of his/her teaching (*facial feedback*). Three facial expressions were used from the ones validated in the previous experiment, and having the highest recognition rate on the valence axis in three of the categories tested previously. The expressions were mapped to contact sensing as a three level discrete feedback as follows: the *happy* expression was used when all three fingers of the robot were in contact with the object, the *content* expression was used when one finger lost contact or the overall contact was weak, and the *annoyed* expression was used when at least two fingers lost contact. The types of feedback described above were provided for the whole duration of the interaction, in all phases of the teaching procedure.

The study addressed two research questions:

RQ1: Does the feedback provided influence the teaching procedure and the learned manipulation model?
RQ2: Does the effect that the type of feedback has on the subjective usability ratings change in relation to task performance?

Participants The participants ($N = 57$, 14 females and 43 males) were selected from university staff and represented the 25–35 years age group. The selection criterion was to not be directly working with robots. Participants were distributed as follows: 12 took the experiment in the first setup (no feedback), 16 were assigned to verbal feedback, 14 to GUI feedback and 15 to facial feedback.

Study Protocol Before beginning the experiment, participants were given general guidelines and were shown a descriptive movie of the teaching procedure. For all setups, the experiment consisted of providing three rounds of demonstration through kinesthetic teaching, of 90 seconds each. Each demonstration round was followed by the robot's replay of the recorded motion. The model learning took place offline after each replay step and was followed by a round of 90 seconds of testing. A post-experiment questionnaire was employed to assess users' satisfaction with the outcome of the teaching task. The total length of the experiment for each participant was 40 to 45 minutes.

Measurements For each round robot measurements consisted of joint angles values for the three fingers used in the task, and the contact signature consisting of tactile response and 3D contact normals. Four objective metrics were computed based on these measures, as defined in [26]:

(1) *range of motion*,
(2) *contact times*,
(3) *joint shakiness* and
(4) *contact error*.

The *range of motion* is based on the difference between the minimum and maximum joint angle values for each finger. These ranges of joint angles are combined in four groups by summing the proximal and distal ranges of motion for thumb, index and middle fingers and separately for the thumb opposition angle. This measure allowed us to compute the percent of the range of motion that was actually demonstrated (when the robot was holding the object) out of the total possible range of motion for a given joint group.

Several metrics have been computed related to *contact times*:

(a) the percent of time two fingers and
(b) three fingers were in contact with the object, out of the total demonstration time; and
(c) the time in force closure, representing the percentage of the total demonstration time in which the three fingers were in contact with the object and the resulting grasp attained force closure [3].

The time in force closure was used as a measure of grasp stability and adaptation quality. The grasping quality was evaluated as described in [21]. *Joint shakiness* represented a measure of the instances of jerky movements. It was evaluated in the testing phase and represented the difference between the raw and smoothed joint velocities averaged across the testing period. *Contact error* represented the difference between the contact value that was predicted (the target) and what the controller executed (the actual) contact value. It gives an overall assessment of the adaptation provided.

Responses from standardized post-experiment questionnaires were used to assess user satisfaction. The questionnaires involved: (1) NASA (Task Load Index) TLX [14], (2) System Usability Scale (SUS) [5] and (3) AttrakDiff [15].

The questionnaires were given in English and clarifications have been provided when necessary. NASA-TLX [14] is commonly used in studies of interface design. It is a workload assessment tool used for evaluating how the user perceived the physical, mental, and temporal demand during a task, and perceived levels of effort, performance and frustration. It consists of six questions, answered with a rating on a 21 point-scale, providing an overall workload score. The SUS questionnaire [5] was used for assessing the overall satisfaction with the system. It consisted of 10 statements

(five positives, five negatives) rated on a five-point Likert scale [18]. Positive questions are given a rank according to the value of their index position minus 1, while negative questions, have a contribution of 5 minus their index position. The score was computed by summing the contribution of each individual component and multiplying the sum by 2.5. AttrakDiff [15] is a method for assessing complementary aspects of the user experience: (1) pragmatic quality, (2) hedonic quality and (3) attractiveness.

However, in this study the hedonic quality of identity was not tested, due to the fact that the robot together with the interface being examined do not represent a commercial application. Thus, the modified version of the questionnaire consisted of 19 pairs of sets of opposite words, which users evaluated on a seven-step scale ranging from −3 to 3. Finally, the participants' assessment of the teaching procedure was evaluated separately by answering four questions considering: how easy was the teaching, how satisfied the participant was with the resulted model, if the robot behaved as expected, and how comfortable the participant felt while providing the demonstrations. Answers were graded on a five point Likert scale.

4.2 Results

Results are presented both with respect to objective, task-specific metrics, as well as subjective user evaluation. Task completion time is constant among users as the teaching and testing rounds were time restricted to 90 seconds.

4.2.1 *Measures of Performance*

All the task-specific metrics presented below are computed based on joint and pressure values of the three fingers that are in contact with the object. They are evaluated for each round of teaching, and represent an evaluation of the learned model. Analysis of variance (ANOVA) was conducted using the measures of performance as dependant variables and the type of feedback as main factor.

Range of Motion The range of motion is a percentage representing how much each joint group moved with respect to the total possible range of motion for that group. The robot's ability to adapt over a higher range of motion shows that a higher range of postures was demonstrated by the user. Detailed statistics are presented in Table 2. The best results were obtained for the verbal feedback setup (E2). Participants that were given graphical or facial display feedback (E3 and E4) explored a significantly lower range of possible motion compared to the case when they were given no feedback at all (E1), as seen in Fig. 7(a). A main effect of the experimental setup was found on the Range of Motion of each finger, see Table 2, last column.

Contact Times The percentage of time when two fingers and three fingers are in contact with the object, out of the total testing time, was evaluated. A high time is an indication of a good adaptation, while a poorly trained motion results in the robot being stiff in that region and losing contact with the object when perturbed. The experimental setup used had

Table 2 Objective Metrics, averaged across all rounds in the testing phase

Objective measures	Experimental setups					
	E1. No Fb	E2. Verbal Fb	E3. GUI Fb	E4. Expr Fb	$F_{\text{statistics}}$	p-value
Range of motion [Deg]						
Thumb Opposition	42.24 ± 26.38	94.50 ± 29.48	44.01 ± 24.36	32.66 ± 17.75	$F(3, 171) = 59.25$	$p < 0.001$
Thumb Finger	29.65 ± 13.85	38.70 ± 07.43	32.20 ± 13.47	27.99 ± 10.99	$F(3, 271) = 7.91$	$p < 0.001$
Index Finger	24.36 ± 10.75	33.81 ± 06.03	33.23 ± 17.81	30.44 ± 14.39	$F(3, 171) = 4.10$	$p = 0.008$
Middle Finger	27.28 ± 13.74	35.59 ± 06.09	34.24 ± 16.27	26.25 ± 13.40	$F(3, 171) = 6.036$	$p = 0.001$
Contact times [% out of Total Time]						
2-Fingers Contact	0.98 ± 0.03	0.99 ± 0.004	0.99 ± 0.003	0.99 ± 0.03	$F(3, 171) = 7.58$	$p < 0.001$
3-Fingers Contact	0.96 ± 0.05	0.97 ± 0.018	0.99 ± 0.01	0.97 ± 0.005	$F(3, 171) = 10.84$	$p < 0.001$
Time in Force Closure	0.67 ± 0.31	0.72 ± 0.15	0.38 ± 0.24	0.58 ± 0.26	$F(3, 171) = 35.159$	$p < 0.001$
Joint shakiness [Deg/s]						
Thumb Opposition	0.029 ± 0.015	0.015 ± 0.004	0.025 ± 0.010	0.023 ± 0.008	$F(3, 171) = 15.091$	$p < 0.001$
Thumb Finger	0.128 ± 0.068	0.080 ± 0.020	0.108 ± 0.045	0.118 ± 0.042	$F(3, 171) = 9.327$	$p < 0.001$
Index Finger	0.110 ± 0.049	0.065 ± 0.021	0.097 ± 0.036	0.097 ± 0.038	$F(3, 171) = 12.779$	$p < 0.001$
Middle Finger	0.093 ± 0.038	0.060 ± 0.016	0.094 ± 0.042	0.098 ± 0.034	$F(3, 171) = 12.835$	$p < 0.001$
Grasping Quality ($\times 10^{-3}$)	0.39 ± 0.04	0.6 ± 0.22	0.22 ± 0.19	0.20 ± 0.12	$F(3, 171) = 23.845$	$p < 0.001$

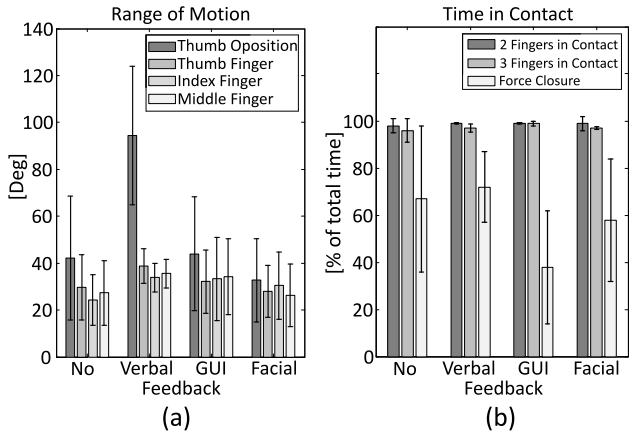

Fig. 7 (**a**) Range of motion and (**b**) Contact Times evaluated for each experimental setup

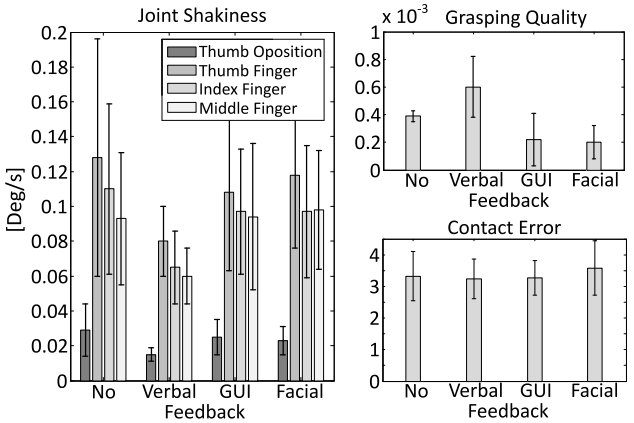

Fig. 8 (**a**) Joint Shakiness, (**b**) Grasping Quality, and (**c**) Contact Error, evaluated for each experimental setup

a significant effect on all the contact times metrics defined, as seen in Fig. 7(b). The percentage of time when two fingers were in contact with the object was lowest when the participant was not given any feedback ($M = 0.98$, $SD = 0.03$, $F(3, 171) = 7.58$, $p < 0.001$) and similarly when three fingers were in contact ($F(3, 171) = 10.84$, $p < 0.001$). However, an important observation is the fact that the percentage of time that three fingers were in contact with the object was highest when the graphical user interface (E3) was used as feedback ($M = 0.99$, $SD = 0.01$), while the second best result was obtained for both the facial display (E4) setup ($M = 0.97$, $SD = 0.005$) and the verbal feedback (E2) setup ($M = 0.97$, $SD = 0.018$). These results together with the negative correlation existing between the time three fingers are in contact and average range of motion (Pearson $r = -0.42$), in the case of E3 and $r = -0.38$ for E4, suggest that while the feedback provided may have been distracting, keeping the user focused on the display rather than on exploring the motion space, helped improve contact accuracy. A decreasing trend was observed for the time in force closure as more feedback was being provided and similarly in the grasping quality, as shown in Fig. 8(b).

Shakiness The Shakiness is also an indication of proper adaptation, with lower values being desirable. The average Range of Motion and average Shakiness are inversely correlated (Pearson $r = -0.58$). Detailed results are presented in Table 2. A significant interaction effect of the experimental setup on the Shakiness values was observed for all joint groups (see Fig. 8(a)) The lowest shakiness values were found in the verbal feedback setup, followed by facial feedback.

Contact Error Contact error decreased considerably as more feedback was provided, as seen in Fig. 8(c), yielding the significant effect ($F(3, 83) = 3.78$, $p = 0.01$) that the experimental setup had on achieving a more stable contact and a smother adaptation. The lowest contact error was achieved when verbal feedback was provided ($M = 3.23$, $SD = 0.62$), while the highest contact error ($M = 3.58$, $SD = 0.85$) is associated with facial feedback.

4.2.2 *Interaction during Demonstration*

User's behavior while providing demonstrations was of particular interest as it would influence the quality of the teaching. We were interested in finding factors that will keep the user engaged in the interaction, in order to ensure good quality demonstrations and also to be willing to provide an optimal number of demonstrations for the robot to be able to properly learn the task. The *demonstration* phase is important for recording proper joint angles. In the *replay* step, the robot will replay the recorded motion while also recording tactile information and thus generating a set of data not influenced by the tutor. For the teacher this step can give a clear understanding of what the robot has recorded (e.g. if the demonstrator moved too fast, only some points in the trajectory will be recorded and this will result in a shaky reproduction). Users' initial attitudes in relation to the feedback being provided influenced the learning by modifying user reaction times, the exploratory motions performed or the observed test patterns, as discussed below.

Exploratory Motions According to our observations, (consistent with the ones mentioned in [13]), in all cases the initial interaction with the robot was driven by users' curiosity. The subjects were not familiar with any humanoid robotic platform, and were not given time to familiarize with our robot before the experiment. However, in the first round of providing demonstrations, they performed a lot of exploratory motions: either pushing the robot to the joints limits, or on the contrary starting from small motions, to try to understand how to control all the degrees of freedom in the robot's fingers. This behavior resulted in frequently lost contact, shaky motions, and an overall poor demonstration.

Losing contact between the robot's fingers and object results in poor replay and thus fewer pairs of postures and contact signatures to be included in the model in the first round of teaching. The improvement rates increase with the rounds of teaching. In several cases, assessing the model improvement across users, regardless of the setup, showed that the improvement rate dropped in the third round of teaching, even if the user was now familiar with the robot capabilities. This might have been due to user fatigue or might be a result of seeing little adaptation while testing the previously obtained models. Exploratory motions performed by the user are necessary in order to get familiar with the robot and to understand the robot's limits. In the case of facial feedback, seeing that the robot was responsive to user actions seemed to encourage subjects to use caution when teaching, which, however, negatively influenced the objective metrics: e.g. the range of motion, see Fig. 7(a).

The subjects were asked to perform a minimum of three demonstrations, but were not limited to an upper number. Interestingly, only three subjects decided to perform a fourth demonstration (two from the verbal feedback setup and one from the GUI setup). Their overall performance ratings during the testing phase were not the best in comparison with other subjects, but they managed to successfully control the robot degrees of freedom so as to teach a wide range of motions.

Model Testing After each learning session the users were asked to test the obtained model in order to decide what should be improved in the next round of teaching. In almost all of the testing cases ($M = 92$ %, $SD = 14.7$), regardless of the experimental setup, the user pushed the robot outside the trained range of motion. During teaching, four types of movements were possible: left and right translational movements, and left and right rotational movements. However, in more than 80 % of the cases, regardless of the feedback provided, in the first round of demonstration only translational movements were trained, but in the testing phase, rotational movements for which no adaptation occurred were tested as well. During the second and third round of demonstration, rotation movements started to be taught, with a higher frequency on the verbal feedback setup. In two cases of users from the E1 setup (no feedback), rotational movements were not trained at all.

User Reaction Times Results showed that human adaptation time was better when either facial feedback or graphical display feedback was being provided. The time between the moments in which the contact was lost and when the human adjusted the fingers positions was lower. The user provided a motion such as to immediately correct the posture. However, this may result in a shaky, sudden motion, thus explaining the high shakiness in these two experimental setups (see Fig. 8(a)). The fastest response time occurred in the case of facial feedback ($M = 1.35$ s, $SD = 0.52$), while the slowest response was recorded for the no-feedback case of ($M = 7.56$ s, $SD = 3.81$).

4.2.3 Subjective Evaluation

For the first two experimental setups (no feedback and verbal feedback) a general interaction assessment was made verbally by the participants. More than 80 % of the participants characterized the interaction as "interesting", "motivating" and "captivating". They also described the shortcomings of the interaction as being the "lack of previous knowledge about the robot" and "the little time available for providing demonstrations". As we were interested in finding the best robot-provided feedback that would improve the interaction, the participants in the other two experimental setups were subject to a more thorough evaluation, being asked to fill in standardized usability questionnaires. Results are presented below.

The effect of experimental setup on task load was not significant. Results (see Table 3) show that mental demand and physical demand were perceived as higher when facial feedback was provided (E4 setup) compared to the case when a graphical display was used (E3 setup). However, the level of frustration perceived was much lower when facial expression feedback was provided and similarly the effort perceived was lower, suggesting that it represents a more natural means of interaction.

The experimental setup did not have a significant effect on the SUS ratings. However, the participants in E4 rated the interface more positively on two key aspects than the users in E3: the usage frequency (namely they would like to use the system more frequently) and the ease of learning the functionality of the system. Results of the SUS questionnaires are summarized in Table 4. An assessment of how attractive the users found the teaching framework was made and results are reported in Table 5. The users taking part in the facially displayed feedback setup (E4) rated the interface higher on hedonic quality and attractiveness, than the

Table 3 NASA Task Load Index (TLX)

	TLX Factors and overall score	
	E3. GUI Fb mean ± std	E4. Expr Fb mean ± std
Mental Load	08.33 ± 04.79	10.18 ± 04.35
Physical Load	04.38 ± 02.32	08.31 ± 06.61
Temporal Load	07.07 ± 03.20	06.43 ± 04.30
Performance	09.84 ± 05.65	09.56 ± 04.70
Effort	09.00 ± 03.46	08.87 ± 04.68
Frustration	08.61 ± 04.94	05.93 ± 05.01
Total Score	47.23 ± 24.36	49.28 ± 29.65

Table 4 System Usability Evaluation (SUS)

	System usability evaluation	
	E3. GUI Fb mean ± std	E4. Expr Fb mean ± std
Usage Frequency	2.93 ± 1.22	3.06 ± 0.85
System Complexity	2.26 ± 0.96	1.81 ± 0.98
Ease of Use	3.20 ± 1.42	3.50 ± 1.26
Technical Support	2.20 ± 1.26	2.31 ± 1.30
Function Integration	3.40 ± 0.82	3.37 ± 0.71
System Inconsistency	2.20 ± 1.01	2.12 ± 1.08
Learn to Use	2.93 ± 1.48	3.25 ± 1.12
Cumbersome	1.86 ± 0.99	1.62 ± 0.80
Confidence	3.06 ± 1.33	3.00 ± 1.03
Previous Knowledge	1.80 ± 1.32	1.81 ± 0.91
Total Score	64.66 ± 15.20	64.68 ± 8.41

Table 5 AttrakDiff Ratings

	AttrakDiff ratings	
	E3. GUI Fb mean ± std	E4. Expr Fb mean ± std
Pragmatic Quality PQ	0.56 ± 0.57	0.40 ± 0.35
Hedonic Quality HQ	0.47 ± 0.43	0.75 ± 0.66
ATT Score	0.74 ± 0.40	1.07 ± 0.64

Table 6 User Evaluation of the Teaching Procedure

	Teaching procedure	
	E3. GUI Fb mean ± std	E4. Expr Fb mean ± std
Ease of Teaching	03.40 ± 01.12	03.12 ± 01.08
Satisfaction	02.80 ± 01.26	03.31 ± 00.60
Expectation	02.86 ± 01.12	03.31 ± 00.94
Comfortability	02.93 ± 01.48	03.50 ± 01.15

users given only graphically displayed feedback (E3). What is more, in the group of words describing the attractiveness, they all assigned the maximum value for the positive attributes ("pleasant", "likeable" "inviting", and "creative"), suggesting that the E4 setup was more motivating and appealing.

Results from evaluating the teaching procedure are presented in Table 6. The participants in the facial feedback experimental setup E4 reported an increased satisfaction with the resulted model ($M = 3.31$, $SD = 0.94$) than those offered only the GUI feedback ($M = 2.86$, $SD = 1.12$), even though the performances in terms of objective metrics were clearly lower. Moreover, the subjects in E4 reported an increased perception of the fact that the robot behaved as they expected. This suggests that seeing a responsive robot increased the users contentment with respect to the interaction. The subjects that took part in E4 reported being significantly more comfortable ($M = 3.5$, $SD = 1.15$) than participants in E3 ($M = 2.93$, $SD = 1.48$), suggesting that facially displayed emotions facilitated a positive interaction.

5 Discussion and Conclusions

This paper addressed the problem of finding a suitable type of feedback that would facilitate robot's learning in a PbD context. Making the human–robot interaction rewarding and keeping the user engaged contributes to improving robot's learning. Two user studies were presented. The first one evaluated the correct classification of 20 robot-expressed facial emotions into given categories. The study targeted testing the assessment that users can relate to robot displayed emotions just as well as they can do with human emotions, and also that they perceive the relative order of emotions, when the valence and arousal levels vary. Results showed that this hypothesis is confirmed only for small levels of granularity, implying fewer emotion categories. For robot-expressed emotions, through LEDs, there was a good recognition rate along Russell's valence axis (differentiating between positive and negative emotions) and a poor recognition rate along the arousal axis. We found little to no support for the second hypothesis: that a high user's self-assessed skill in recognizing human emotions might positively impact the ability to recognize robot expressions. The third hypothesis, that intense emotions take a very small reaction time, was supported.

Limitations of this study are threefold.

First, the LED display used for generating the facial expressions could not portray a good enough range of human emotions. We aimed to determine a small set of best recognized facial displays, however, the displayed faces raised problems in terms of ambiguity of the LED display. The expression of the same emotion might look different when using another robotic platform.

Second, the lack of prior interaction with the robot or its expressions made the respondents unsure when assigning extreme intensity values for the displayed emotions without having a prior idea of the possible range.

Third, we did not assess participants' ability to recognize the same facial expression when displayed by a human face.

The second user study was conducted to assess the usability of a teaching by demonstration interface that was not initially based on a user-centered design. In our approach, similar to other robot teaching tasks, the interaction was initiated by the human. We designed the interaction in a way that would ease teaching for the human user, by having rounds

of demonstration, robot replay and testing. This allowed not only the iterative refinement of the obtained model, but it also helped the user to understand what the robot has learned at each step and what needs to be improved in the next demonstration. Different feedback modalities were used to reflect the strength of the contact between the robot's fingers and the object: verbal feedback, graphical user interface feedback, facial displayed feedback and no feedback at all.

Results presented confirmed that the type of feedback provided by the robot influences both subjective and objective metrics. According to objective metrics, satisfactory results were obtained in all study cases. During testing, three fingers are in contact with the object in more than 95 % of the time, force closure grasps are attained for more than one third of the testing time, and no large differences can be seen between shakiness and grasping quality across setups. While in most cases the verbal feedback from a knowledgable person proved the best, this is not feasible in real world applications.

According to the subjective metrics evaluation, the experimental setup influenced the ease of the interaction, user demand and friendliness. Participants no longer saw the interaction as restricting and no longer perceived a temporal pressure when being provided feedback in a natural manner. While the verbal feedback yielded the best results, this made the user dependent on an external expert, present at all times. The GUI feedback required technical knowledge and understanding of the mapping of touch sensors to the displayed interface, while the facial expressions feedback proved to be very intuitive and stimulating. For a naive user who is not familiar with the robot this may be the best way of obtaining a satisfactory model for manipulating an object in a comfortable and rewarding interaction.

The feedback provided by the robot is similar to the social cues that humans might use when teaching another person, have the advantage of giving the user an intuitive understanding of the robot's limits. This might well compensate for a lack of prior knowledge, while keeping the user focused and motivated in the interaction for longer periods of time. This makes the proposed method appropriate for novice users.

Future work in the direction of using social cues in PbD should address the question of what is the optimum level of feedback that should be provided to the user. Particularly in our experiment, mapping facial displays to how strong the contact on the fingertips was had a great impact on improving the time the fingers were in contact. However, this was not enough for our task success since the task also required exploring the range of motion. Therefore, mapping the range of motion to another social cue, such as voice or hand gestures done with the other hand, might have increased the task success rate even further.

Subsequently task performance could be improved by assessing how the type of feedback influences the users' approach of the task. Namely providing feedback to systematically guide the users' training and testing could lead to an improved robot performance.

Acknowledgements The research leading to these results has received funding from the Swiss National Science Foundation through the NCCR in Robotics, the European Community's Seventh Framework Program FP7/2007-2013—Challenge 2—Cognitive Systems, Interaction, Robotics—under grant agreement no[231500]-[ROBOSKIN], and the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement no 288533 ROBOHOW.COG.

References

1. Bartneck C, Reichenbach J, Breemen AV (2004) In your face, robot! The influence of a character's embodiment on how users perceive its emotional expressions. In: Design and emotion 2004 conference
2. Bartneck C, Kanda T, Mubin O, Mahmud AA (2009) Does the design of a robot influence its animacy and perceived intelligence? Int J Soc Robot 1(2):195–204
3. Bicchi A (1995) On the closure properties of robotic grasping. Int J Robot Res 14(4):319–334
4. Breazeal C (2009) Role of expressive behaviour for robots that learn from people. Philos Trans R Soc Lond B, Biol Sci 364:3527–3538
5. Brooke J (1996) SUS—a quick and dirty usability scale. In: Usability evaluation in industry, pp 189–194
6. Cakmak M, Thomaz AL (2012) Designing robot learners that ask good questions. In: HRI, pp 17–24
7. Calinon S, Billard A (2007) What is the teacher's role in robot programming by demonstration?—Toward benchmarks for improved learning. Interact Stud 8(3):441–464, Special issue on psychological benchmarks in human–robot interaction
8. Cramer H, Goddijn J, Wielinga B, Evers V (2010) Effects of (in)accurate empathy and situational valence on attitudes towards robots. In: Proceedings of the 5th ACM/IEEE international conference on human–robot interaction, pp 141–142
9. Dautenhahn K (1998) The art of designing socially intelligent agents—science, fiction and the human in the loop. Appl Artif Intell J 12:12–17 special issue on socially intelligent agents
10. Dautenhahn K, Werry I (2000) Issues of robot-human interaction dynamics in the rehabilitation of children with autism
11. Gielniak MJ, Thomaz AL (2011) Spatiotemporal correspondence as a metric for human-like robot motion. In: Proceedings of the 6th international conference on human–robot interaction, pp 77–84
12. Giusti L, Marti P (2006) Interpretative dynamics in human robot interaction. In: Robot and human interactive communication, RO-MAN, pp 111–116
13. Hanson D (2005) Expanding the aesthetics possibilities for humanlike robots. In: Proc IEEE "Humanoid robotics" conference. special session on the Uncanny Valley
14. Hart S, Staveland L (1988) Development of NASA-tlx (task load index): results of empirical and theoretical research. In: Human mental workload, pp 139–183
15. Hassenzahl M, Burmester M, Koller F (2003) Attrakdiff: Ein fragebogen zur messung wahrgenommener hedonischer und pragmatischer qualität. In: Mensch & computer 2003: interaktion in bewegung, vol 196. B.G. Teubner, Leipzig, p 187
16. Knox WB, Glass BD, Love BC, Maddox WT, Stone P (2012) How humans teach agents—a new experimental perspective. Int J Soc Robot 4(4):409–421

17. Leyzberg D, Avrunin E, Liu J, Scassellati B (2011) Robots that express emotion elicit better human teaching. In: Proceedings of the 6th international conference on human–robot interaction, HRI '11, pp 347–354
18. Likert R (1932) A technique for the measurement of attitudes. Arch Psychol 22(140):1–55
19. Nicolescu MN, Mataric MJ (2003) Natural methods for robot task learning: instructive demonstrations, generalization and practice. In: Proceedings of the second international joint conference on autonomous agents and multiagent systems, pp 241–248
20. Peacock M (2001) Match or mismatch? Learning styles and teaching styles in efl. Int J Appl Linguist 11(1):1–20
21. Ponce J, Sullivan S, Sudsang A, daniel Boissonnat J, Merlet JP (1996) On computing four-finger equilibrium and force-closure grasps of polyhedral objects. Int J Robot Res 16:11–35
22. Robins B, Dautenhahn K, Nehaniv CL, Mirza NA, Francois D, Olsson L (2005) Sustaining interaction dynamics and engagement in dyadic child-robot interaction kinesics: lessons learnt from an exploratory study. In: IEEE international workshop on ROMAN 2005, pp 716–722
23. Russell JA (1980) A circumplex model of affect. J Pers Soc Psychol 39(6):1161–1198
24. Russell J (1991) Culture and the categorization of emotions. Psychol Bull 110(3):426–450
25. Russell J (1997) Reading emotions from and into faces: resurrecting a dimensional-contextual perspective. In: Russell JA, Fernández-Dols JM (eds) The psychology of facial expression. Cambridge University Press, Cambridge, pp 295–320
26. Sauser EL, Argall BD, Metta G, Billard AG (2012) Iterative learning of grasp adaptation through human corrections. Robot Auton Syst 60:55–71
27. Ushida H (2010) Effect of social robot's behavior in collaborative learning. In: Proceedings of the 5th ACM/IEEE international conference on human–robot interaction, HRI '10, pp 195–196
28. Yohanan S, MacLean KE (2011) Design and assessment of the haptic creature's affect display. In: Proceedings of the 6th international conference on human–robot interaction, HRI '11, pp 473–480

Ana Lucia Pais is a Ph.D. student in the Learning Algorithms and Systems Laboratory (LASA) at the Swiss Federal Institute of Technology in Lausanne (EPFL). She received her Bachelors and Masters degree (2011) from the Polytechnical University of Bucharest (UPB), Romania. Her research interests focus on using PbD techniques for improving HRI and task learning.

Brenna D. Argall is an Assistant Professor of Electrical Engineering and Computer Science at Northwestern University. She holds a Faculty Research Scientist position at the Rehabilitation Institute of Chicago, where she is founder and director of a laboratory for rehabilitation robotics research. Prior to joining NU and RIC, she was a postdoctoral fellow (2009–2011) at the EPFL. She received in 2009 her Ph.D. from the Robotics Institute at Carnegie Mellon University, where she also completed in 2006 a M.S. in Robotics and in 2002 a B.S. in Mathematics. Her research focuses on robotics, machine learning and rehabilitation.

Aude G. Billard is Associate Professor and head of the LASA Laboratory at the School of Engineering at the EPFL. Prior to this, she was Research Assistant Professor at the Department of Computer Sciences at the University of Southern California, where she retains an adjunct faculty position to this day. Aude Billard received a B.Sc. (1994) and M.Sc. (1995) in Physics from EPFL, with specialization in Particle Physics at the European Center for Nuclear Research (CERN), an M.Sc. in Knowledge-based Systems (1996) and a Ph.D. in Artificial Intelligence (1998) from the Department of Artificial Intelligence at the University of Edinburgh. Her research interests focus on machine learning tools to support robot learning through human guidance. This extends also to research on complementary topics, including machine vision and its use in human–machine interaction and computational neuroscience to develop models of learning in humans.