Exploring the use of a handheld device in language teaching human-robot interaction

Pierre Rouanet and Pierre-Yves Oudeyer
FLOWERS team - INRIA Bordeaux Sud-Ouest
http://flowers.inria.fr

Abstract

In this paper, we are exploring a human-robot interface allowing non-expert users to easily and intuitively teach new words to a robot. Many challenges may be addressed before achieving such behavior as joint attention, naming, categorization and searching. Instead of using direct interactions (such as gesture and voice recognition) which are not robust enough in unconstrained environments, we use here a handheld device as a mediator between the user and the robot. The device allows us, among other things, to display the robot's camera visual feedback as well as interacting through user's gestures on the touchscreen. Thus, users are able to draw the robot's attention toward locations, to select a particular object in the field of view of the robot, to name it and with our "active-searching" system to find it again later. An exploratory study has been carried out in order to get some early users opinions, which tend to show that users easily use the interface during robust interaction sessions.

HRI challenges associated with language teaching

Social robots are drawing an increasing amount of interest both in scientific and economic communities. These robots should typically be able to interact naturally and intuitively with non-expert humans, in the context of domestic services or entertainment. Yet, an important challenge needs to be addressed: providing robots with the capability to adapt and operate in uncontrolled, novel and/or changing environments, in particular when interacting with non-expert humans.

Among the many issues raised by such types of social robots, we focus here on the problem of how a non-expert human can teach a new word to a robot, typically associated with an object in the environment. In its full generality, this brings up very hard problems, in particular the issue of how a robot can infer the conceptual meaning of a new word [18]. Here, we will restrict ourselves to the case where a given word is only associated with a single concrete object: thus, we are not dealing with concepts, but only with visual appearance recognition. Nevertheless, this is a very ambitious project since several important obstacles still need to be crossed:

Attention drawing : How can a human smoothly, robustly and intuitively draw the attention of a robot towards himself and towards the interaction when the robot is doing its own activity?

Pointing and joint attention : How can a human draw the attention of a robot towards the object he wants to show to this robot? If the object is not in the field of view of the robot, how to push the robot to move adequately? When the object is within the field of view, how can the human point at this particular object? How can the human understand what the robot is paying attention to? How can joint attention be realized [10, 2]?

Naming : How can the human introduce a symbolic form that the robot can perceive, register, associate with the object, and later on recognize when repeated by the human? What modaliti(es) shall be used to ensure ease of use, naturalness, and robustness?

Categorization : How can associations between

words and visual representations of objects be memorized and reused later on to allow the human to have the robot search an object associated with a word he has already taught to the robot? Like when human children learn language, social partners can only try to guide the acquisition of meanings but cannot program directly the appropriate representations in the learner's brain. Thus, the process of data collection may lead to inappropriate learning examples. False interpretations could ensue from a wrong data collection. How can we maximize the efficiency of example collection while keeping intuitive and pleasant interaction with non-expert humans?

Searching : How can a human intuitively ask for the robot to find an already known word? How can easily and robustly the matching word can be recognized? How can the user intuitively improve the recognition rate ?

One could try to address several of these challenges by transposing human-human modes of interaction based on gesture recognition, gaze tracking or voice recognition [15][16][19]. In principle, this approach would provide really natural interactions. Unfortunately, existing associated techniques are not robust enough in uncontrolled environments (noise, lighting, occlusion...) and most social robots have a body whose shape and perceptual apparatus is not compatible with those modes of interaction (small angle of view, small height...). This implies that such an approach is bound to fail if one is interested in intuitive and robust interaction with non-expert users in unconstrained environments.

Objectives

In the project outlined in this abstract, we did not focus on the machine learning challenge but on the HRI challenge. We argue that one way to help to achieve intuitively and robustly some of the functionalities presented above is to develop simple artefacts that will serve as mediators between the human and the robot to enable natural communication, in much the same way as icon based board-based artefacts were developed for leveraging natural linguistic communication between human and certain bonobos [14]. More particularly, we argue that using mobile devices, such as illustrated in figure 1 may enable to circumvent some of the above mentioned problems. Though it may seem less natural to use a device as a mediator between humans and robots, by allowing a robust, reliable and working interaction without consideration of environmental constraints it may lead to actually more practical and usable interactions. Such interfaces may provide pleasant and nonrestrictive interactions and so rather quickly become sort of "natural" interactions.

Figure 1: Teaching new words to a robot with the help of a handheld device

Kemp et al. have shown how a laser pointer can be intuitively used to designate objects to a robot [12]. As we try to do in our system, they can draw the robot attention toward objects and so realize joint attention between the human and the robot. However their robot is able to automatically grasp the object from the detected 3D spot, in a framework that requires image segmentation algorithm and/or a priori objects knowledge. If objects are not known beforehand these are still hard problems. In order to circumvent these problems, we argue in this paper, that is possible to have the user segmenting himself the object from the image in an intuitive manner by using a handheld touchscreen device as a mediator. Indeed, the screen of the device can be used to provide the human with information about what the robot is perceiving, but also to transfer information from the human,

through easily perceivable gestures, to the robot [13]. In particular, here we can display the camera stream on the screen and ask the user to circle on the touch-screen the interesting object. Moreover, handheld devices allows the human to be next to the robot and physically engaged, for example allowing to catch object and waving them physically in the robot's field of view. Finally, they also allow for tele-interaction with the robot through the video feedback of the camera.

Contrary to tangible user interfaces (TUI) such as Guo's [8], we argue here that PDA-based interface are differently perceived. Indeed, they imply different metaphors that lead to different users' expectations. Users perceived TUI as remote controls and so expect for immediate responses. On the other hand, PDA is probably here more seen as a more "high-level" interface. Furthermore, monitoring the screen takes attention away from the robot, and so users seem to assume more autonomy in robot's behavior and expect to be able to send more complicate commands. In spite of these assets, a handheld device also implies a small screen: this constrains the kinds of interfaces one can design and leveraging it to increase mutual human-robot comprehension [6] is an interesting challenge.

Many PDA-based interface have been developed over the last few years. Some are inspired by classical human-computer interfaces, using components such as menus and widgets [4][9][11]. In this paper, we make the hypothesis that this kind of interfaces are not best-suited to social robotics interactions. Indeed, they are not intuitive interactions, particularly for non-expert users, and so often require a training period. Moreover, we assume here that non-entertaining interfaces can have negative impacts on the user's experience during the interaction.

Calinon et al. developed another PDA-based interaction allowing to improve gesture and speech recognition by using the camera and the micro of a PDA [3]. This interface allows to teach associations between user's gestures and user's verbal utterance. The PDA is embedded into the robot, so staying close enough of the robot is needed in order to be able to teach it something. Furthermore, as previously noticed they assume to have a good image segmentation algorithm.

Fong developed a tele-operation driving system for a mobile robot by using a virtual keyboard. He also used metaphors such as point-and-click and used the sketches-inputs on the robot visual feedback in order to design paths [7]. We try here to extend this approach to other scopes of applications such as teaching new words to a robot.

In order to design a user-centered interface [1] which suits users expectations and allows to teach new words efficiently, we follow the "design-implementation-user studies" cycle. In this paper, we report on the first iteration of this cycle, which is thus yet highly exploratory.

Outline of the system

An exploratory version of the system was designed to be tested with a mobile robot such as the AIBO robot, and based on the use of a Pocket PC. Here, we chose to use hand drawn words as the way words are given as input by the human. Another possibility would have been to use a virtual keyboard on the Pocket PC or speech voice. Hand drawn words were used because they are easier to perceive and better recognized than speech and yet potentially faster and more pleasant to write than using a virtual keyboard.

In this system, the screen of the handheld device displays the video stream of the robot's camera (about 15 fps). It accurately shows what the robot is looking at, which can thus be monitored by the user allowing to resolve the ambiguity of what the robot is really seeing.

When the human wants to show to the robot an object which is not in its field of view, the user can sketch on the screen to make it move in an appropriate position: vertical strokes for forward/backward movements and horizontal strokes for right/left turns. Elementary heuristics are used to recognize these straight sketches. The moves of the robot are continuous until the user re-touch the screen in order to stop it. Another stroke can directly be drawn to go on the next move (for instance, go forward then directly turn right). Pointing on a particular point on the screen makes the robot look at the corresponding spot. The tilt and pan value of the head are computed using the camera field of view constants in order to accurately center the robot's sight on the chosen location.

When the user wants to show an object which is in the field of view of the robot, and thus on the

screen, it sketches a circle around this object on the touch screen (see figure 2). As for the straight strokes, heuristics are here used to recognize circular sketches, based on the shape of the stroke and the distance between the first and the last point of the sketch. Circling is a really intuitive gesture because users directly "select" what they want to draw attention to. Moreover, this gesture is particularly well-suited to touch-screen based interactions. Schmalstieg et al. used the circling metaphor to select objects in a virtual world [17]. Circling is also a crucial help for the robot since it provides a rough visual segmentation of the object, which is otherwise a very hard task in unconstrained environments. With the stroke and the background image, we can extract the selected area and define it as our object's image. A classical geometry algorithm is used to test the belonging to the polygon formed by the stroke.

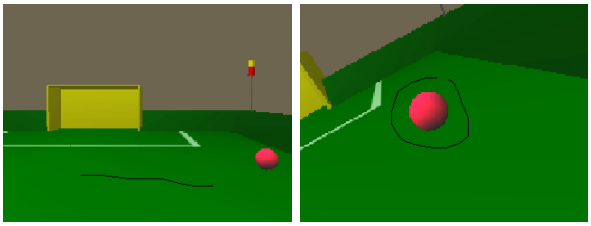

Figure 2: Drawing attention towards an object: the user first sketches directions to position the robot such that the object is in its field of view (left), and then encircles what he wants to show to the robot (right).

Once this object is encircled, a menu pops-up showing several interaction options. The "name" choice makes the system wait for a referent (symbol, word, sketch...) drawn on the screen (see figure 3). Once the user has sketched the word he wanted to teach, the association between the hand drawn word sketch and the visual features inside the circle is memorized. The word is recorded as a regularly time spaced 2D points list along the user's stroke. It allows fast and robust distance matching measures with a dynamic time warping algorithm. The extracted image is then associated into our data structure. Later on, when the user directly sketches a word without first encircling an object, a matching distance measures is done on every recorded stroke and the nearest neighboor is found. This nearest neighboor is called "recognized word" (which is a sketch and not a list of symbols). The robot understands that the user would like it to search for the image of the object associated to the recognized word. Standard color histogram distances are used to track the object image on the video stream of the camera with the OpenCV library [1]. A simple search algorithm have been developed to move the robot until it detects the searched object.

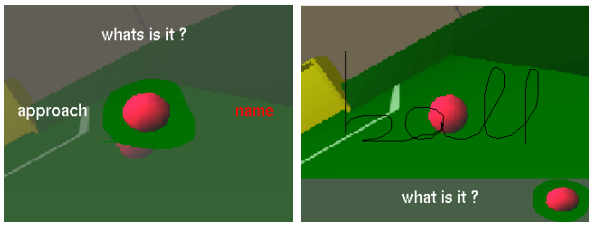

Figure 3: When an object has been shown to the robot, the user can decide to provide a name/word for it, inputted as a graphical sketch (which can be the written version of the word or any other drawing).

Thus, the categorization/detection system is here based on a simple memory-based algorithm. This could easily be improved in the future. Yet, its robustness can be improved thanks to a feature of the interface which we call "active searching".

Due to hand-writing variability, recognition can be erroneous. To improve it, a small set (about 3) of close enough words are selected by their sketches' distances. In order to easily be identified by users, their associated visual representations (the image of the object) are presented on the screen as shown in figure 4. Then, the user chooses the desired one by "clicking" it on the screen, and the robot goes to search for it. An analogy can be developed between our strategy, which we call "active searching" and search engines on the internet. Indeed, presenting a small set of images to the user can be compared with the results list in a search engine request. The small set of images represents here the top of the request results, among which the user chooses what he find interesting. Always choosing

[1] http://opencv.willowgarage.com/wiki/

the closest word, as most classical matching algorithms do, may be analogous to to always use the "I'm feeling lucky" button in google. Providing to the user a small set of choice allows to drastically improve results as opposed to having the robot directly search for the closest match. Moreover, this allows users to better realize how the robot comprehends words and their similarities with each other.

Figure 4: When the handwriting recognition does not produce a reliable result, a small set of possible matches are shown to the user to allow him to decide which is the good one.

Exploratory study

Methods

While this interface is already working with real AIBO robots, initial user-studies were conducted using the AIBO Webots simulator for practical reasons. On top of that, using a virtual test environment allows us to use simple object recognition algorithm (only based on color histogram) with quite good results. In a real environment with change of lighting, much robust algorithms should have been implemented.

The participants of the user-study were ten students, (10 male, 0 female) and aged from 20 to 28 (Mean: 23) recruited on the University of Bordeaux. They used a Sam-sung UM-PC as handheld device and interact with a stylus. They described themself as unfamiliar with robot or Tablet PC. The aim of this study was to get primary user feedbacks and pilot our next studies.

Before the beginning of the experiment, the tester was quickly introduced to the scope of application. Then a user guide, with a complete interface capacities description, was given to them and finally their experiment goals were described.

Three goals were defined:

- Get the robot to look at three particular objects placed in the virtual environment.
- Teach the robot a name for each of them.
- Once the labeled objects are out of view ask the robot to search for one (figure 5).

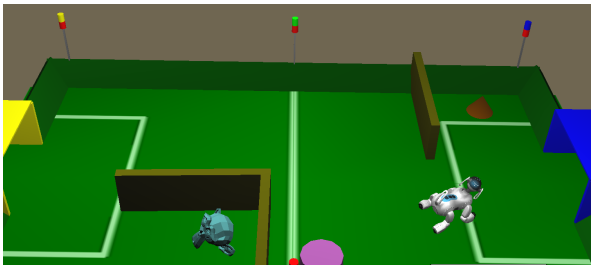

Figure 5: The test environment

Afterwards, they can learn to use the interface during a short time (about one minute), then they had 10 minutes to achieve their objectives. All testers achieved their goals before the time limit. Once they finished, they had to answer a questionnaire mainly made up questions about using the interface. All the answers were given on a Likert scale from "strongly disagree" (-2) to "strongly agree" (2).

- How easily understanding the use of the interface was ?
- How efficient the interface was ?
- How ease-of-use the interface was ?
- How entertaining using the interface was ?

There also was more specific questions about the handwriting system and the active searching system :

- How easy handwriting on the screen was ?
- How pleasant handwriting on the screen was ?

- Would you prefer another system such as virtual keyboard or voice recognition ? Which one ?
- How easily understood the use of the active searching system was ?
- How efficient the active searching system was ?
- How pleasant using the active searching system was ?

Results

The small group of participants and the absence of a real robot do not allow to get very reliable results. Here, they are most considered as indications for future developments and experiments.

8 on 10 participants stated they feel they have learned quickly or really quickly how to use the whole system. Driving the robot appears to cause most problem, only 6 on 10 testers felt they have learned quickly how to move the robot. 8 on 10 participants stated the interface was efficient enough in this context. Only 1 on 10 stated that the interface was not efficient to drive the robot. 7 on 10 stated the interface was really easy to use and 3 on 10 stated it was rather easy to use. 6 on 10 participants stated the interface was neither entertaining nor restrictive. Only 4 on 10 stated it was entertaining.

Only 5 on 10 participants found the handwriting system easy or really easy to use, while 3 on 10 found it difficult or really difficult. 6 on 10 testers stated they found the handwriting neither pleasant nor unpleasant. 3 on 10 found it rather pleasant and 2 found it rather unpleasant. Only 1 on 10 participants stated he prefers to use a virtual keyboard when 6 on 10 stated they would not. On the other hand 7 on 10 thought they prefer to use a vocal commands system.

Finally 9 on 10 participants found the active searching system easy or really easy to understand, when only 1 found it neither easy nor hard to understand. Similar results were found for the efficiency of the active searching system. 8 on 10 found it pleasant while 2 found it neither pleasant nor unpleasant.

Discussion

The motion system of the AIBO robot was the main problem encountered by the participants during the tests. After informal interviews, it appears that users tend to assume intelligence in robot's moves, for instance, they expected the robot to automatically stop in front of walls and to avoid obstacles. They also claimed for more high-level displacements, like the ability to draw paths on the touchscreen. As we argue before, the PDA interface is considered by the users as a high-level interface and so lead to specific user's expectations.

Evaluating only the interface was a major problem here because users tend to focus more on the robot abilities and recognition algorithms than the system of interaction. In order to get more interesting and reliable results, improvements in these directions should be done. The bounciness of perceived images inherent with legged robots, seems to also have a negative impact on the efficiency of the interaction.

Although users managed to teach new words to the robot, they asked for a voice recognition system arguing it would be better suited to the situation.

Active searching was really naturally accepted, easily used and positively perceived.

Figure 6: In our new prototype, the iPhone accelerometer can be used as a TUI to orient the robot head.

Further work

Further work will span several dimensions. First, new driving interactions must be provided, such as paths mentioned above. Second, an exploratory study with real robots is needed, both with legged robots and wheeled robots in order to accurately evaluate the bounciness impact on the efficiency of the interaction. Third, diverse modes of word inputting, including through the virtual keyboard and through speech, must be compared. Fourth, we will go further on the active searching analogy. Indeed, we can easily display the next set of results (if the "good one" wasn't in the first "page"). Fifth, we will investigate the use of multi-touch handheld devices, which are bound to provide intuitive and richer ways of interacting with the robot. We already have a prototype of this system running on the iPhone. New driving methods have been developed using the accelerometer to steer the robot, or to orient the head in a particular direction (see figure 6). The multi-touch gesture recognition allows different motion type: for instance, one finger gestures for accurate motion (they stop as the user touches up the screen) and two fingers gestures for continuous motions. Some informal trials seem to identify this new system is both more efficient, more robust, and entertaining than our initial system. Sixth, we will introduce visual representation and recognition algorithm based on SIFT texture features and visual bag-of-words recognition [5] to improve their robustness, particularly in real environments. Finally, we need to refine our evaluation methodology, which includes the ability to separate features of the interaction system and features of the robot itself, as well as to separate the evaluation of the different functionalities of the system (moving to appropriate position, attention drawing, naming, searching).

References

[1] Julie A. Adams. Critical considerations for human-robot interface development. In *AAAI Fall Symposium on Human-Robot Interaction*, Cape Cod, MA, November 2002.

[2] Cynthia Breazeal and Brian Scassellati. Infant-like social interactions between a robot and a human caregiver. *Adapt. Behav.*, 8(1):49–74, 2000.

[3] S. Calinon. Pda interface for humanoid robots using speech and vision processing. 2003.

[4] Dalgalarrondo, Dufourd, and Filliat. Controlling the autonomy of a reconnaissance robot. In *SPIE Defense & Security 2004 Symposium. Unmanned Ground Vehicle Technology VI Conference*, 2004.

[5] D. Filliat. A visual bag of words method for interactive qualitative localization and mapping. In *Proceedings of the International Conference on Robotics and Automation (ICRA)*, 2007.

[6] Terrence Fong, Nathalie Cabrol, Charles Thorpe, and Charles Baur. A personal user interface for collaborative human-robot exploration. In *6th International Symposium on Artificial Intelligence, Robotics, and Automation in Space (iSAIRAS)*, Montreal, Canada, June 2001.

[7] Terrence W Fong, Chuck Thorpe, and Betty Glass. Pdadriver: A handheld system for remote driving. In *IEEE International Conference on Advanced Robotics 2003*. IEEE, July 2003.

[8] Cheng Guo and Ehud Sharlin. Exploring the use of tangible user interfaces for human-robot interaction: a comparative study. In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 121–130, New York, NY, USA, 2008. ACM.

[9] Helge Huttenrauch and Mikael Norman. Pocketcero – mobile interfaces for service robots. In *In Proceedings of the Mobile HCI, International Workshop on Human Computer Interaction with Mobile Devices*, 2001.

[10] F. Kaplan and V. Hafner. The challenges of joint attention. *Proceedings of the 4th International Workshop on Epigenetic Robotics*, 2004.

[11] Hande Kaymaz Keskinpala Julie A. Adams Kazuhiko Kawamura. Pda-based human-robotic interface. In *Proceedings of the IEEE International Conference on Systems, Man & Cybernetics: The Hague, Netherlands, 10-13 October 2004*, 2003.

[12] Charles C. Kemp, Cressel D. Anderson, Hai Nguyen, Alexander J. Trevor, and Zhe Xu. A point-and-click interface for the real world: laser designation of objects for mobile manipulation. In *HRI '08: Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, pages 241–248, New York, NY, USA, 2008. ACM.

[13] Marjorie Skubic Sam Blisard Andy Carle Pascal Matsakis. Hand-drawn maps for robot navigation. In *AAAI Spring Symposium, Sketch Understanding Session, March, 2002.*, 2002.

[14] Sue S. Rumbaugh and Roger Lewin. *Kanzi : The Ape at the Brink of the Human Mind.* Wiley, September 1996.

[15] A. Haasch S. Hohenner S. Huwel M. Kleinehagenbrock S. Lang I. Toptsis G. Fink J. Fritsch B. Wrede G. Sagerer. Biron – the bielefeld robot companion. In *Proc. Int. Workshop on Advances in Service Robotics Stuttgart Germany 2004 pp. 27–32.*, 2004.

[16] Brian Scassellati. Mechanisms of shared attention for a humanoid robot. In *Embodied Cognition and Action: Papers from the 1996 AAAI Fall Symposium*, 1996.

[17] Dieter Schmalstieg, Luis Miguel Encarnação, and Zsolt Szalavári. Using transparent props for interaction with the virtual table. In *I3D '99: Proceedings of the 1999 symposium on Interactive 3D graphics*, pages 147–153, New York, NY, USA, 1999. ACM.

[18] Luc Steels and Frederic Kaplan. Aibo's first words: The social learning of language and meaning. *Evolution of Communication*, 4(1):3–32, 2000.

[19] Kai Nickel Rainer Stiefelhagen. Real-time recognition of 3d-pointing gestures for human-machine-interaction. In *International Workshop on Human-Computer Interaction HCI 2004, May 2004, Prague (in conjunction with ECCV 2004)*, 2004.