A Study of three Interfaces allowing Non-expert Users to Teach New Visual Objects to a Robot and their Impact on Learning Efficiency

Pierre Rouanet, Pierre-Yves Oudeyer, David Filliat

HAL Id: inria-00541794
https://hal.inria.fr/inria-00541794

Submitted on 1 Dec 2010

A Study of three Interfaces allowing Non-expert Users to Teach New Visual Objects to a Robot and their Impact on Learning Efficiency

Pierre Rouanet and Pierre-Yves Oudeyer*
Flowers Team INRIA Bordeaux Sud-Ouest
pierre.rouanet@inria.fr

David Filliat**
UEI - ENSTA ParisTech
david.filliat@ensta.fr

Abstract—We developed three interfaces to allow non-expert users to teach name for new visual objects and compare them through user's studies in term of learning efficiency.

Index Terms—Human-Robot interaction, interfaces, joint attention, learning, social robotics

I. INTRODUCTION

Social robots are drawing an increasing amount of interest but to allow these robots to interact naturally and intuitively with humans we need to provide the robot with the capability to operate in uncontrolled and changing environments. We focus here on the problem of how a robot can learn through the interactions with the human and in particular, how a non-expert human can teach a new word, typically associated with a single concrete object in its close environment, to a robot. Several obstacles need to be crossed to achieve such an ability:

- **Attention drawing:** How can a human robustly and intuitively draw the attention of a robot towards himself?
- **Pointing and joint attention:** How can a human designate an object to a robot and draw its attention toward this particular object? 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 could the object be robustly extracted from its background? How can the human understand what the robot is paying attention to? How can joint attention be realized [1][2]?
- **Naming:** How can the human introduce a symbolic form that the robot can perceive and associate with the object, and later on recognize when repeated by the human?
- **Categorization and searching:** 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. How can we maximize the efficiency of example collection while keeping intuitive and pleasant interaction with non-expert humans?

Thus, we have to address *visual recognition, machine learning* and also *Human-Robot Interaction (HRI)* challenges. We argue that, while using state-of-the art incremental machine learning and computer vision algorithms [3], we can by focusing on the HRI challenges significantly improve the whole learning system by allowing the user to provide the robot with good learning examples [4].

II. OUTLINE OF THE SYSTEM

We adopted the "bags of visual words" approach [5] to process images in our system and developed an incremental version suited to HRI [3][4].

Based on this visual perception and machine learning system, we developed three interfaces to provide the user with the following abilities: move the robot, draw its attention toward a direction or an object, define an area within its field of view as a new learning example, and finally associate a word to this visual object.

1) iPhone: This interface used an iPhone as a touch-screen based interface where we display the image perceived by the camera of the robot to allows users to monitor what the robot really sees, which is a key feature to achieve joint attention. The touch-screen is also used to sketch trajectories as motion commands. The user can also encircle an object directly on the screen (fig. 1) to define the selected area as the new learning example. It also provides a rough visual segmentation of the object, which is otherwise a very hard task in unconstrained environments [4]. Then, users can enter a name by using a virtual keyboard or by vocally naming the object.
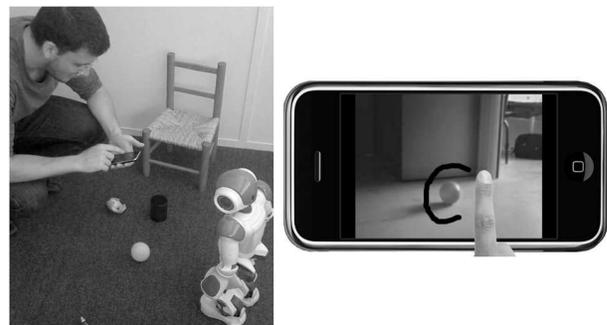

Fig. 1. Encircling an object allows the user to notify the robot that he wants to teach a name for this object and also provides an useful rough segmentation.

2) Wiimote: In this interface, we use the Wiimote accelerometers to map their values to the robot movements as in a classical tangible user interface. We can move both the body and the head of the robot. When the user names the object, as this interface does not provide any way to define the object area, the whole image perceived at the time the word is pronounced, is used as a new learning example.

3) Wiimote and laser: In this interface the user used both a Wiimote and a laser in his hands to teach new words to the robot. The Wiimote is used to move the robot and the laser is used to draw the robot attention toward objects. The robot's head automatically tracks the laser pointer. When the user wants to teach a name for an object that the robot is looking at, he can encircle the object with the laser pointer.

III. EXPERIMENTS

We designed two studies where participants had to teach different objects to the Nao robot and compare the three developed interfaces and in particular their impact on the learning examples provided by users.

A. First experiment : expert vs non-expert users

Here, we study the impact of the feedback of what the robot sees and compare this among experts and non-expert users. Participants had to provide the robot with three learning examples for each five objects located in a very simple environment (13 participants: 5 experts and 8 non-experts).
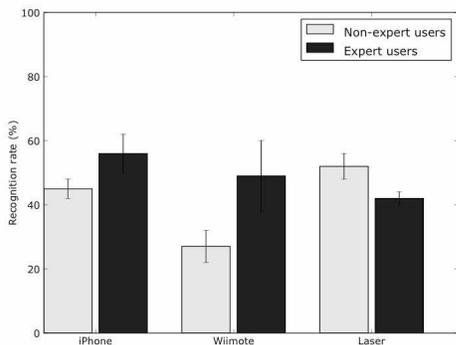

Fig. 2. The iPhone and laser interfaces seem to allow both non-expert and expert users to provide good learning examples, while the Wiimote interface does not allow non-expert users to provide good examples due to the lack of visual feedback. Indeed, they have difficulties to correctly estimate the robot's field of view and so in most learning examples objects were not even in the image.

As shown on the figure 2, the categorization performance are high for the iPhone and laser interfaces. This means that both expert and non-expert users managed to provide good learning examples with these interfaces. While with the Wiimote interface, we can clearly see that the non-expert users did not manage to provide good learning examples. On the other hand, the recognition rate of the expert users is high, showing that this interface is usable, only if the users are able to correctly estimate the field of view of the robot.

B. Second experiment : feedback and encircling

Here, we study the different impact among the various type of visual feedback and the role of encircling. Indeed, with the iPhone interface, the user can exactly monitor what the robot is seeing, while with the laser interface he can only know if the robot is detecting the laser. Furthermore, as encircling

is only useful in a real and thus complex environment, we created a more complex environment. Indeed, it allows us to circumvent the issue of the segmentation of an image, which is very difficult and an ill-defined problem in a general context but becomes trivial with an uniform background.

Here, the participants were only "expert" users, allowing to collect a much larger database than in the first experiment (7 objects, 50 learning examples per object). This introduces a bias, but nevertheless this first experiment showed that no significant differences were measured when experts and non-experts used the iPhone or the Wiimote and laser interfaces.
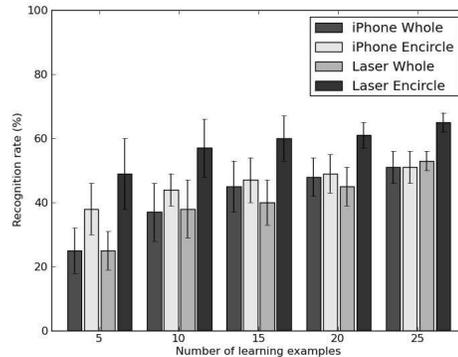

Fig. 3. We can notice that the examples provided by the users, who were encircling the objects on the iPhone, are significantly better than with the other conditions.

As shown on the figure 3, we can notice that there is no significant difference of recognition rate between the two conditions where we do not consider the encircled images. While encircling with the laser improves the learning with few examples it seems to be useless with more examples. On the other hand, the examples provided by encircling with the iPhone lead to a significant higher recognition rate.

Encircling with the laser seems to be not as efficient as encircling with the iPhone. Indeed, the detection of the laser leads to technical issues, such as occlusions, false detections and deformation due to the projection of the detected points in the plane of the camera of the robot. Thus, some of the learning examples provided with the laser were irrelevant or only partially relevant.

REFERENCES

[1] C. Breazeal and B. Scassellati, "Infant-like social interactions between a robot and a human caregiver," *Adapt. Behav.*, vol. 8, no. 1, pp. 49–74, 2000.
[2] F. Kaplan and V. Hafner, "The challenges of joint attention," *Proceedings of the 4th International Workshop on Epigenetic Robotics*, 2004.
[3] D. Filliat, "Interactive learning of visual topological navigation," in *Proceedings of the 2008 IEEE International Conference on Intelligent Robots and Systems (IROS 2008)*, 2008.
[4] P. Rouanet, P.-Y. Oudeyer, and D. Filliat, "An integrated system for teaching new visually grounded words to a robot for non-expert users using a mobile device," in *Proceedings of the Humanoids 2009 Conference*, 2009.
[5] G. Csurka, C. Dance, L. Fan, J. Williamowski, and C. Bray, "Visual categorization with bags of keypoints," in *ECCV04 workshop on Statistical Learning in Computer Vision*, 2004, pp. 59–74.