

An integrated system for teaching new visually grounded words to a robot for non-expert users using a mobile device

Pierre Rouanet and Pierre-Yves Oudeyer*

Flowers Team
INRIA Bordeaux Sud-Ouest
351, cours de la Libération
33405 TALENCE - France
pierre.rouanet@inria.fr

David Filliat**

UEI - ENSTA ParisTech
32 boulevard Victor
75015 PARIS - France
david.filliat@ensta.fr

Abstract—In this paper, we present a system allowing non-expert users to teach new words to their robot. In opposition to most of existing works in this area which focus on the associated visual perception and machine learning challenges, we choose to focus on the HRI challenges with the aim to show that it may improve the learning quality. We argue that by using mediator objects and in particular a handheld device, we can develop a human-robot interface which is not only intuitive and entertaining but will also “help” the user to provide “good” learning examples to the robot and thus will improve the efficiency of the whole learning system. The perceptual and machine learning parts of this system rely on an incremental version of visual bag-of-words. We also propose a system called ASMAT that makes it possible for the robot to incrementally build a model of a novel unknown object by simultaneously modelling and tracking it. We report experiments demonstrating the fast acquisition of robust object models using this approach.

I. INTRODUCTION

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, particularly in the context of domestic services or entertainment. Yet, an important challenge needs to be addressed: providing the robot with the capability to adapt and operate in uncontrolled, novel and/or changing environments, in particular when interacting with non-expert humans. Thus, the robot should have the ability to discover its environment. Among the various challenges that this implies, 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 to a robot typically associated with an object in its close environment.

However, in its full generality, this brings up very hard problems and in particular the issue of how the robot can infer the conceptual meaning of a new word [1]. Here, we will restrict ourselves to the specific case where a given word is only associated with a single concrete object. Thus, we are not dealing with concepts, but only with visual object instance 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? How can the human draw the robot’s attention even if he is not in its field of view?

- **Pointing:** 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 can the human point at this particular object and how could the object be robustly extracted from its background?
- **Joint attention:** How can the human understand what the robot is paying attention to? How can the human accurately know what the robot is seeing? How can joint attention be realized [2][3]?
- **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? How can we recognize when two examples are related to the same object? Could the human help the robot during this process?
- **Searching:** How can a human intuitively ask for the robot to find or point to an already known object? How can easily and robustly the matching word can be recognized? How can the user intuitively help the recognition?

Thus, in order to give the ability to a non-expert human to

teach new words to its robot, we have to address *visual recognition*, *machine learning* and also *Human-Robot Interaction (HRI)* challenges. In this paper, we argue that by focusing on the HRI challenges we could significantly improve the whole learning system. We therefore propose a system to tackle some of these challenges (attention drawing, pointing, joint attention and naming) and illustrate the improvement in the efficiency of the learning system brought by this approach.

II. RELATED WORK

Over the past few years, some works tried to address these issues by transposing human-human modes of interaction. Scassellati developed mechanism of shared attention through gaze monitoring and pointing gesture [4]. In his work, he used a fixed upper-torso humanoid robot. Many researches also tried to recognize pointing gestures in order to designate objects to a robot [5][6]. Steels et al. developed a complete social framework based on direct interactions (pointing gestures and speech recognition) to allow an user to teach words to an AIBO [1]. In this work, the authors are more specifically focusing on the visual perception and machine learning issues. Yet, the HRI was identified has a major limitation of their system. In particular, they showed that the lack of robustness of the interface lead to some bad learning examples and so decreased the learning performance [7]. Thus, in this paper we are proposing an integrated system allowing to teach new words to a robot, with a special attention on the HRI challenges. More specifically, we tried to develop an intuitive, efficient and entertaining interface, which also makes it possible for the user to provide the system with good learning examples. By doing so, we are hoping to improve the performance of the whole learning system.

As presented above, in most of the related works, authors choose to use the direct interactions (gesture recognition, gaze tracking or voice recognition) to try to address the above mentioned HRI challenges. In particular, this approach potentially provides really natural interactions which is particularly important with non-expert users. Unfortunately, existing associated techniques are not robust enough in uncontrolled environments (due to noise, lighting or 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 will fail if one is interested in intuitive and robust interaction with non-expert users in unconstrained environments.

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 artefacts were developed for leveraging natural linguistic communication between human and certain bonobos [8]. More particularly, we argue that using mobile devices, such as illustrated in figure 1 may enable to circumvent some of these problems. Though it may seem less natural to use a device as a mediator



Fig. 1. Using a handheld device as a mediator object to control the movements of a social robot.

object between humans and robots, by allowing a robust, reliable and working interaction, 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.

These kinds of interfaces have already been used to interact with a robot. Kemp. et al have shown how a laser pointer can be intuitively used by people with severe motor impairments to robustly designate objects to a robot [9]. Thanks to the laser spot light, the human can also accurately know what he is pointing at. Yanco et al. used an interface based on an input device (touch screen or joystick) to select objects which will be grasped by a wheelchair mounted robotic arm [10]. In their work the user can directly monitor the object selection on the screen of the device. As we try to do in our system, they both 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 a 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 this problem, 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 touch-screen device. Indeed, the screen of the device can be used to provide the human with information about what the robot is perceiving, which is interesting with non-expert users who are particularly prone to make assumptions about the capacity and behavior of the robot. But it also allows to transfer information from the human to the robot, through easily perceivable gestures [11]. Moreover, these sketches and gestures are natural cues and so are natural for people to use [12]. Thus, we can develop intuitive collaborative interaction allowing the human to supervise the robot and allowing the robot to take advantage from the human capabilities [13]. In particular, here we can display the camera stream on the screen and let the user to encircle the interesting object on the touch-screen. Finally, handheld devices allow the human

to be next to the robot and physically engaged, for example allowing to catch objects and waving them physically in the robot's field of view. They also allow tele-interaction with the robot through the video feedback of the camera.

Other handheld device based interfaces have been developed recently. For instance, Fong et al. used a PDA for remote driving [14], and Kaymaz et al. used it to tele-operate a mobile robot [15]. Sakamoto et al. showed how they can control a house cleaning robot through sketches on a Tablet PC [16]. However, to our knowledge nobody used a handheld device for such teaching interactions. We already proposed a prototype based on a handheld device to teach new words to a robot [17]. This prototype was developed with a special care to the classical design lessons in HRI and HCI [18][19][20]. The exploratory study indicated that it was a promising approach, providing an intuitive and efficient interface for non-expert users. We also compared different interfaces for showing object to a robot and concluded that the gesture interface based on a handheld device was stated as the most satisfying by the users [21]. In this paper, we propose a full-system with advanced visual perception, machine learning and HRI components.

III. OUTLINE OF THE SYSTEM

A. Visual perception

We adopted the popular “bags of visual words” approach to process images in our system. Bags of visual words is a method developed for image categorization [22] that relies on a representation of images as a set of unordered elementary visual features (the words) taken from a dictionary (or code book). Using a given dictionary, a classifier is simply based on the frequencies of the words in an image, thus ignoring any global image structure. The term “bag of words” refers to document classification techniques that inspired these approaches where documents are considered as unordered sets of words. Several applications also exist for robotics, notably for navigation (e.g. [23], [24]).

The words used in image processing are based on local image features such as the SURF keypoints [25] we are using in this paper. As these features can be noisy and are represented in high dimension spaces, they are categorized using vector quantization techniques (such as k-means) and the output clusters of this categorization are the words of the dictionary. Instead of building the dictionary off-line on an image database as is performed in most applications, we use an incremental dictionary construction ([26]) that makes it possible to start with an empty dictionary and build it as the robot discovers its surroundings. This make it possible to learn objects without any a priori on the object type or the environment of the robot.

This model has two interesting characteristics for our application: the use of feature sets make it robust to partial object occlusions and the feature space quantization bring robustness to image noise linked to camera noise or varying illumination. More over, with the incremental dictionary construction, this quantization is adapted as the environment

evolve (for example when light change from natural to artificial).

B. Machine learning

For our application, the classifier designed for object recognition should be trained incrementally, i.e. it should be able to process new examples and learn new objects without the need to reprocess all the previous data. To achieve that, we use a generative method in which training entails updating a statistical model of objects, and classifying entails evaluating the likelihood of each object given a new image.

More specifically, we use a voting method based on the statistics of visual words appearance in each object. The recorded statistics (according to the learning method described later) are the number of appearance a_{wo} of each visual word w of the dictionary in the training examples of each object o . For object detection in a new image, we extract all the visual words from this image and make each word w vote for all objects o for which $a_{wo} \neq 0$. The vote is performed using the *term frequency-inverted document frequency (tf-idf)* weighting [27] in order to penalize the more common visual words. An object is recognized if the quality of the vote result (measured as the difference between the best vote and the second best) is other a threshold.

Estimating the statistics a_{wo} require the labelling of examples with the associated object name. As will be described later, we will use two methods for example labelling depending on the information given by the user : labelling the whole image or labelling only an image area given by the user and representing the object . Additionally, we will propose a new scheme for automatically labelling new examples from an initial user labelled example (see section III-D).

C. Human Robot Interaction

We choose to use the Nao robot as our test platform. Indeed, to us it well represents the present of social robotics: with a toy-aspect and classical inputs (camera, microphone). Furthermore, it is a biped robot and it has a complex skeletal so it leads to complex motions. Finally, as it is a humanoid, user will probably be more prone to teach it new words.

Our system was embedded on an Apple iPhone used as a mediator object between the human and the robot. We choose this device because it allows to display information on the screen to the user and also allows to interact through “natural” gestures through a large amount of possibilities due to the multi-touch capacities. Moreover, thanks to the large success of the iPhone we can take advantage of a well-known interface, allowing ease of use.

In this system, the screen of the handheld device displays the video stream of the robot's camera (at 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 (see figure 2). As mentioned above, achieving such an ability with direct interaction is difficult with personal robots such as Nao humanoid or the AIBO robot due to the specific morphology of such robots



Fig. 2. We display the video stream of the camera of the robot on the screen. This allows to accurately monitor what the robot is seeing and thus really achieving joint attention situation.

and the particular characteristic of their camera, in particular with non-expert users who are specifically prone to make really ambitious assumptions about the robot capacity. Moreover, having a visual feedback seems to entertain the user [21], while the robot is moving for instance.



Fig. 3. 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 if he wants to center the robot's sight on a specific spot just past on it (right).

When the human wants to draw the robot attention toward 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 (see figure 3).

When the user wants to show an object which is in the field of view of the robot, and thus on the screen, in order to teach a name for this object, it sketches a circle around this object on the touch screen (as shown on figure 4). Circling is a really intuitive gesture because users directly “select” what they want to draw attention to. This gesture is particularly well-suited to touch-screen based interactions. For instance, Schmalstieg et al. used the circling metaphor to select objects in a virtual world [28]. 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. This simple gestures has two important functions:



Fig. 4. Encircling an object allows the user to notify the robot that he wants to teach a name for this object. But it also provides an useful rough object segmentation.

- First, it allows to clearly separate, by using two different gestures, the action of drawing the robot attention toward an object and the user's will of teaching a new word for this object.
- Second, 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. Classical computer graphics algorithms are used to compute this area (Bresenham line drawing and flood fill).



Fig. 5. Some objects can not be segmented with the classical object segmentation algorithms. For instance, on the left the object has almost the same color than the background. In the middle example, the object is not movable and so can not be segmented with the motion based segmentation method. On the right example, an automatic method can not guess if the user wants to show only the head of the giraffe or the whole poster. Furthermore, this is a 2D object and so the range method can not deal with it.

In this paper, we argue that object segmentation is still a hard task in unconstrained and unknown environments. Different approaches have been developed over the past years to address this problem. However, they are still suffering from a lack of robustness. For example, *Region growing algorithms* try to address this problem by determining regions where color or texture are homogeneous. These region are iteratively expanded from a seed [29]. Yet, these algorithms can not deal with complex objects made up of several sub-parts with various colors and textures. Moreover the colors or textures of the object can also be similar to the background

(examples are shown on the figure 5). A lot of researches have also studied how the boundaries of the object could be determined through its motion : *motion based segmentation* [30]. Although, this algorithm can only segment carryable or movable object. *Range segmentation* algorithms use images containing depth information to compute regions belonging to the same surface [31]. Of course, this approach does not allow to recognize 2D objects such as a poster. By asking the user to segment the image with a circling stroke, we circumvent all the above mentioned problems and we can deal with all the kind of objects, allowing us to work in unconstrained environments.

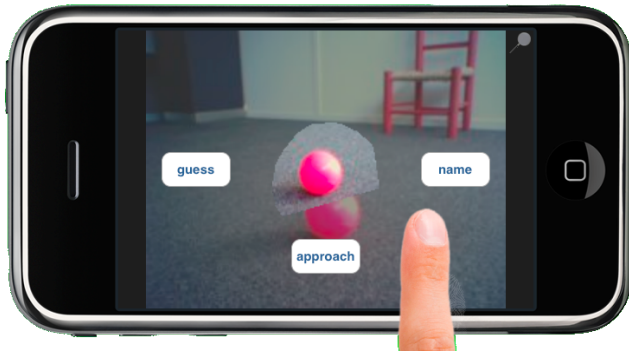


Fig. 6. When a joint attention situation has been achieved on an object, a set of actions are presented to the user, which can in particular decide to teach a name for this object

Once this object is encircled, a menu pops-up showing several interaction options : *naming*, *guessing* or high-level actions such as *approach*, as drawn on figure 6. The “name” choice makes the system wait for the user to enter a referent. In this prototype, we choose to enter the name through a virtual keyboard which allows us to quickly have a whole working system (see figure 7). Moreover, it also allows to circumvent the name recognition problem. Obviously, in further work other modalities such as vocally naming will be used. Once the user has entered the word he wanted to teach, the visual features inside the circle are added to our learning system and the corresponding words are labelled as belonging to the model of the object (as described in section III-B).



Fig. 7. When the user has decided to teach a name for an object, the system is waiting for him to enter a word with the virtual keyboard.

Later on, when the robot has learnt some words, the user can ask it to look for one of this object by entering the looking menu and selecting the object he wants the robot to look for, as shown on figure 8. A simple search algorithm

have been developed to move the robot until it detects the searched object by using the visual bag-of-words system.

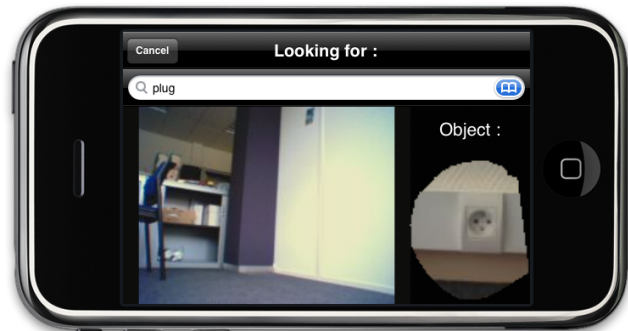


Fig. 8. Once the robot has learnt some words, the user can ask the robot to look for one, by selecting it on the looking menu.

D. Active simultaneous modelling and tracking : ASMAT

As presented above, each time the user encircles an object, our system only improves the model of the object with a single image. However, the appearance of an object can drastically changes from different points of view, so multiple images of an object are needed in order to be able to recognize it, in spite of the point of view [32]. So we develop a system allowing to automatically collect many learning examples, with different points of view, from a single user’s intervention (see figure 9). Through this approach, we think we can get a much more robust model of an object and provide a less restrictive interaction to the users.



Fig. 9. We can automatically extract image of an object from different points-of-view with our *active simultaneous modelling and tracking (ASMAT)* system in which the robot turns around the objects, and thus improve the model of the object and make it more robust.

With this system, when the user encircles an object on the screen of the device, we make the robot turn around the object and take pictures of the object from different points of view. However, this a hard problem because we do not have any a priori model of the object in order to track it, and to get a model of the object we need to be able to track it. In order to achieve such an ability we need to be able to simultaneously model and track an object without any prior data. A related system was already presented by Dowson et al. [33], but using previously recorded video, while we can here directly define the movements of the camera. For this reason, we called our system *active simultaneous modelling and tracking (ASMAT)*. With this system, the robot incrementally constructs a model of the object, thanks to the

incremental bag-of-words method which is used at the same time to robustly track the object and thus to enable the robot to turn around it.

The robot incrementally turns around the object through N steps. The robot goes from one position to the next one by moving sideways and turning itself in order to keep the tracked object in the center of its sight. For each stationary position, we lightly move the head of the robot to quickly get M different images of the object. For each image, we first find the SURF keypoints (in order to work in real time) matching the model of the object. We compute the gravity center of these points and compute the average distance to this center. We then filter the points which are too far from this average region. Finally, we define the bounding box of the elected points as the object, i.e. we add to the model of the object all the keypoints inside this box (see algorithm 1).

Such an approach allows the learning system to quickly get a robust model of the object (as shown in the experiments in section IV-B). However, this system can lead to exponential deviation due to fact that the constructed model is also used to track the object, thus the more the system is mistaken more he will be mistaken. To circumvent this problem, we could display on the screen of the iPhone the bounding box of the tracked object allowing the user to stop the robot as soon as it goes wrong and for instance, ask the user to re-encircle the object in order to restart the process. However, this possibility has not yet been implemented and thus evaluated.

Algorithm 1 ASMAT(*user_encircled_image*)

keypoints \leftarrow *extract_keypoints*(*user_encircled_image*)
update_object_model(*keypoints*)

while not *user_stop*() **and** $i < N$ **do**

for j **in** 1 **to** M **do**

move_robot_head()

keypoints \leftarrow *extract_keypoints*(*robot_camera*)

matches \leftarrow *find_matching_object_model*(*keypoints*)

elected \leftarrow *filter_isolate_points*(*matches*)

bb \leftarrow *compute_bounding_box_from_points*(*elected*)

for each kp **in** *keypoints* **inside** *bb* **do**

update_object_model(kp)

end for

end for

walk_step_around_object()

keypoints \leftarrow *extract_keypoints*(*robot_camera*)

matches \leftarrow *find_matching_object_model*(*keypoints*)

center \leftarrow *compute_gravity_center*(*matches*)

robot_center_sight(*center*)

$i \leftarrow i + 1$

end while

IV. EXPERIMENTS

A. Encircling

In order to test our integrated recognition system (visual perception, machine learning and HRI), we needed to recre-

ate a realistic test environment. First of all, we characterize the environment and the kind of object the humans would teach to robot in a home environment. To us, these objects will probably be everyday objects which can be found for instance in a living room. We can then define two main groups of objects:

- small, light and carryable objects as a newspaper, keys or a ball
- bigger, fixed objects as a plant or a plug

We think that the first categories will represent the most important part of taught objects. Furthermore, they also probably are the hardest to recognize due to the background changes. To our knowledge, the most matching database would be the ETH-80 image set. However, this database uses class object and not only instances. Furthermore, a neutral background is used which is not representative of unconstrained environments. The point of view is also quite similar from one image to another, while the robot will have to recognize an object from different points of view. Thus, we decided to create our own database with a special attention to our criteria. We chose 20 different everyday objects, which are rather small, carryable and well-textured. As our objects are carryable, we must be able to recognize them in spite of their location, i.e. with different backgrounds. So we chose five different backgrounds (on the ground, on a desk, at a window, in front of a bookcase and in the kitchen) and took two pictures by object and by background. So, finally we got 10 images by object taken with different points of view but at the rather same height (figure 10). Every images has been roughly segmented with a stroke encircling the object. Our database is deliberately rather small because we want to be able to recognize an object with few learning examples provided by the user. To us, ten user's examples seems to already be a maximum in order to keep an nonrestrictive interaction. The image were taken and encircled by the authors. Furthermore, they were taken with the camera of the iPhone, and converted to a low resolution (320x240) to correspond to a common resolution.

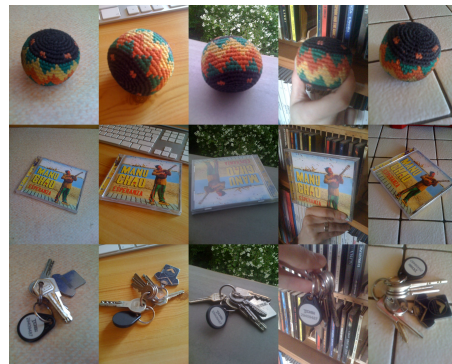


Fig. 10. Several examples of a specially constructed objects database corresponding to the kind of objects we think an human would like to teach to its robot (everyday object, small and carryable). The images were taken with different backgrounds in order to be able to recognize objects in spite of their location.

In order to test our recognition system, we use the following experimental protocol.

- We randomly choose N images per object.
- We train our learning system with these images.
- We test our learning system with the other images ($10 - N$ images per object).
- The test is repeated 20 times in order to circumvent the randomize effect.
- The final results are the mean recognition rate of each test.

As shown on the figure 11, encircling the objects allows us to improve the recognition rate by 20% in average. So, we can see that we the recognition rate maximum (about 80%) was reached by giving six encircled learning examples, while the maximum with nine full images was not reached. By simply encircling the objects on the screen, the user can improve our recognition system and in particular can achieve robust recognition with fewer learning examples. Thus, we can reduce the number of user’s interventions.

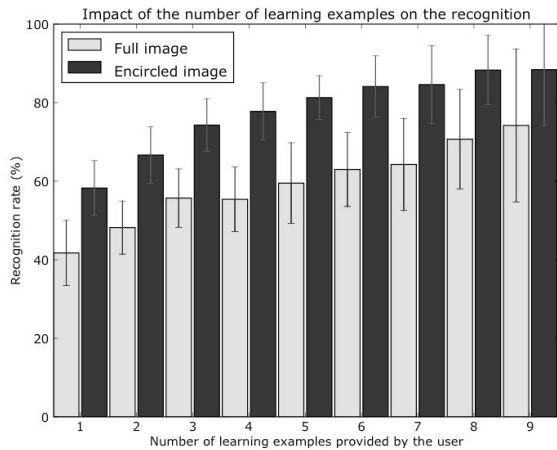


Fig. 11. We can notice that with the encircled images, we have a recognition rate superior of 20% in average than with the full image. Thus, we reach the maximum recognition rate faster (from the sixth encircled example with encircled image while it is still not reached with nine full learning examples).

However, our images were here not gathered in “real” conditions. Indeed, the images were not taken with a robot by non-expert users. Nevertheless, we try to reproduce a plausible interaction, by taking pictures with possible angles of view, with a common resolution for camera... Furthermore, the encircling was done by the authors. Thus, we should, in a future experience, test if non-expert users would provide as good inputs, as were given by the expert users and see if we can get the same results.

B. ASMAT

As mentioned above, we developed a system to automatically get a larger set of learning examples by making the robot turn around the objects. We try to evaluate the impact of such a method on the learning process. So, we designed an experiment where a user taught four different objects to the

robot by encircling them on the screen of an iPhone. Then, the robot automatically turned around the object. At each step, the robot moves sideways and forward. Then it turns in order to recenter its sight on the center of the tracked objects. Then we take five snapshots with lightly different positions of the head. We repeat this operation five times, so 25 images were taken by learning example. For each object, the user give five different learning examples with different points of view of the object. We define two conditions:

- In condition A, we only use the first image (the one encircled by the user) to train our recognition system.
- In condition B, we use all the 25 images labelled using our ASMAT system to train the recognition system.

We then used a similar database as the one used above (5 backgrounds, 2 images per background and per object), with our four objects to test the quality of the learning.

As we can see on the figure 12, with the condition A, we have a linear progression of the recognition rate according to the number of learning examples : with five learning examples we reached about 60% of recognition. With the condition B, we can notice a really fast increase of the recognition rate. A maximum (about 80%) is reached from the second learning examples. We can also notice that this maximum is not reached with the condition A even after the fifth examples. Thus, the ASMAT system seems to allow the getting of a robust and reliable model of an object with really few user’s interventions.

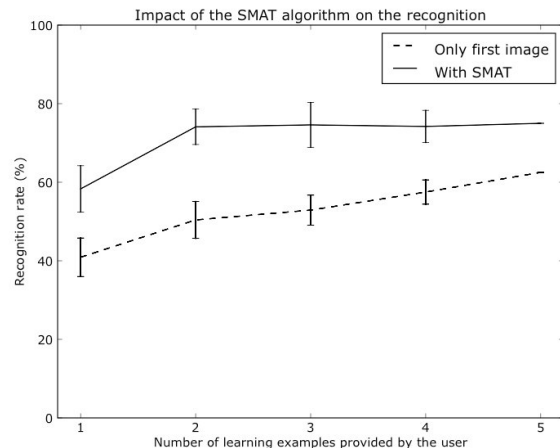


Fig. 12. Recognition rate according to the number of examples given by the user. We define two conditions, one with the ASMAT, the other without it. We can notice that by using this system, we can get a more accurate model of an object with fewer user’s learning examples.

V. CONCLUSION AND FUTURE WORKS

A. Conclusion

Our proposed system, based on a combination of advanced HRI, visual perception and machine learning methods, allows non-expert users to intuitively and robustly teach new words to their robot. By using the touch-screen to transfer information, we have developed collaborative interactions, improving

the mutual understanding between the robot and the human. We also showed that the interface may help the user to provide good learning examples which will thus improve the whole learning system.

B. Future works

In future works we will try to evaluate the impact of different interfaces on the learning by designing a comparative user's study with different kind of interfaces (with an iPhone, with a laser pointer and with direct interactions). We will compare them with a "learning quality" measure, but also with satisfaction questionnaires to assess their usability. It would also be interesting to evaluate the impact of the ASMAT system on the user's experience, especially with non-expert users. Thus, it could enhance the interaction, by making it more lively and more entertaining for the users. On the other hand, the extra time taken to do the active learning, may fatigue users.

Then, we will use a vocal naming system and so we will have to be able to compare two vocal words. We will also allow the user to improve the learning through collaborative interactions, where he could help the clustering of the different learning examples, and try to evaluate the real impact of such a feature.

VI. ACKNOWLEDGMENTS

The authors would like to thank Jérôme Béchu for his implication in the development and the realization of the different parts of the system.

REFERENCES

- [1] L. Steels and F. Kaplan, "Aibo's first words: The social learning of language and meaning," *Evolution of Communication*, vol. 4, no. 1, pp. 3–32, 2000. [Online]. Available: <http://www3.isrl.uiuc.edu/junwang4/langev/localcopy/pdf/steels02aiboFirst.pdf>
- [2] 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.
- [3] F. Kaplan and V. Hafner, "The challenges of joint attention," *Proceedings of the 4th International Workshop on Epigenetic Robotics*, 2004.
- [4] B. Scassellati, "Mechanisms of shared attention for a humanoid robot," in *Embodied Cognition and Action: Papers from the 1996 AAAI Fall Symposium*, 1996.
- [5] K. Nickel and R. Stiefelwagen, "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.
- [6] V. V. Hafner and F. Kaplan, "Learning to interpret pointing gestures: Experiments with four-legged autonomous robots," in *Proceedings of the KI2004 Workshop on Neurobotics*. Springer, 2004, pp. 225–234.
- [7] F. Kaplan, *Les machines apprivoisées comprendre les robots de loisir*. vuibert, 2005.
- [8] S. S. Rumbaugh and R. Lewin, *Kanzi : The Ape at the Brink of the Human Mind*. Wiley, September 1996.
- [9] Y. S. Choi, C. D. Anderson, J. D. Glass, and C. C. Kemp, "Laser pointers and a touch screen: intuitive interfaces for autonomous mobile manipulation for the motor impaired," in *Assets '08: Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility*. New York, NY, USA: ACM, 2008, pp. 225–232.
- [10] K. Tsui, H. Yanco, D. Kontak, and L. Beliveau, "Development and evaluation of a flexible interface for a wheelchair mounted robotic arm," in *HRI '08: Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*. New York, NY, USA: ACM, 2008, pp. 105–112.
- [11] M. Skubic, S. Blisard, A. Carle, and P. Matsakis, "Hand-drawn maps for robot navigation," in *AAAI Spring Symposium, Sketch Understanding Session, March, 2002.*, 2002.
- [12] M. Goodrich and J. Olsen, D.R., "Seven principles of efficient human robot interaction," *Systems, Man and Cybernetics, 2003. IEEE International Conference on*, vol. 4, pp. 3942–3948 vol.4, Oct. 2003.
- [13] T. Fong, N. Cabrol, C. Thorpe, and C. 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. [Online]. Available: citeseer.ist.psu.edu/fong01personal.html
- [14] T. W. Fong, C. Thorpe, and B. Glass, "Pdadriver: A handheld system for remote driving," in *IEEE International Conference on Advanced Robotics 2003*. IEEE, July 2003.
- [15] H. Kaymaz, K. Julie, A. Adams, and K. 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.
- [16] D. Sakamoto, K. Honda, M. Inami, and T. Igarashi, "Sketch and run: a stroke-based interface for home robots," in *CHI '09: Proceedings of the 27th international conference on Human factors in computing systems*. New York, NY, USA: ACM, 2009, pp. 197–200.
- [17] P. Rouanet and P.-Y. Oudeyer, "Exploring the use of a handheld device in language teaching human-robot interaction," in *Proceedings of the AISB 2009 Workshop : New Frontiers in Human-Robot Interaction*, 2009.
- [18] J. A. Adams, "Critical considerations for human-robot interface development," in *AAAI Fall Symposium on Human-Robot Interaction*, Cape Cod, MA, November 2002.
- [19] J. Drury, J. Scholtz, and H. Yanco, "Awareness in human-robot interactions," *Systems, Man and Cybernetics, 2003. IEEE International Conference on*, vol. 1, pp. 912–918 vol.1, Oct. 2003.
- [20] H. A. Yanco, J. L. Drury, and J. Scholtz, "Beyond usability evaluation: analysis of human-robot interaction at a major robotics competition," *Hum.-Comput. Interact.*, vol. 19, no. 1, pp. 117–149, 2004.
- [21] P. Rouanet, J. Béchu, and P.-Y. Oudeyer, "A comparison of three interfaces using handheld devices to intuitively drive and show objects to a social robot : the impact of underlying metaphors," *RO-MAN*, 2009.
- [22] 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.
- [23] J. Wang, R. Cipolla, and H. Zha, "Vision-based global localization using a visual vocabulary," in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA)*, 2005.
- [24] A. Angeli, D. Filliat, S. Doncieux, and J.-A. Meyer, "Real-time visual loop-closure detection," in *Proceedings of the International Conference on Robotics and Automation (ICRA)*, 2008.
- [25] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Comput. Vis. Image Underst.*, vol. 110, no. 3, pp. 346–359, 2008.
- [26] D. Filliat, "Interactive learning of visual topological navigation," in *Proceedings of the 2008 IEEE International Conference on Intelligent Robots and Systems (IROS 2008)*, 2008.
- [27] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *IEEE International Conference on Computer Vision (ICCV)*, 2003.
- [28] D. Schmalstieg, L. M. Encarnação, and Z. Szalavári, "Using transparent props for interaction with the virtual table," in *I3D '99: Proceedings of the 1999 symposium on Interactive 3D graphics*. New York, NY, USA: ACM, 1999, pp. 147–153.
- [29] R. Adams and L. Bischof, "Seeded region growing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 6, pp. 641–647, 1994.
- [30] A. Arsenio, P. Fitzpatrick, C. C. Kemp, and G. Metta, "The whole world in your hand: Active and interactive segmentation," pp. 49–56, 2003. [Online]. Available: <http://cogprints.org/3329/>
- [31] A. Bab-Hadiashar and N. Gheissari, "Range image segmentation using surface selection criterion," *Image Processing, IEEE Transactions on*, vol. 15, no. 7, pp. 2006–2018, July 2006.
- [32] P. Robbel, "Exploiting object dynamics for recognition and control," Ph.D. dissertation, Massachusetts Institute of Technology. Dept. of Architecture. Program in Media Arts and Sciences., 2007.
- [33] N. D. H. Dowson and R. Bowden, "Simultaneous modeling and tracking (smat) of feature sets," *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, vol. 2, pp. 99–105, 2005.