The Impact of Human–Robot Interfaces on the Learning of Visual Objects

Pierre Rouanet, Pierre-Yves Oudeyer, *Member, IEEE*, Fabien Danieau, and David Filliat, *Member, IEEE*

***Abstract*—This paper studies the impact of interfaces, allowing nonexpert users to efficiently and intuitively teach a robot to recognize new visual objects. We present challenges that need to be addressed for real-world deployment of robots capable of learning new visual objects in interaction with everyday users. We argue that in addition to robust machine learning and computer vision methods, well-designed interfaces are crucial for learning efficiency. In particular, we argue that interfaces can be key in helping nonexpert users to collect good learning examples and, thus, improve the performance of the overall learning system. Then, we present four alternative human–robot interfaces: Three are based on the use of a mediating artifact (smartphone, wiimote, wiimote and laser), and one is based on natural human gestures (with a *Wizard-of-Oz* recognition system). These interfaces mainly vary in the kind of feedback provided to the user, allowing him to understand more or less easily what the robot is perceiving and, thus, guide his way of providing training examples differently. We then evaluate the impact of these interfaces, in terms of learning efficiency, usability, and user's experience, through a real world and large-scale user study. In this experiment, we asked participants to teach a robot 12 different new visual objects in the context of a robotic game. This game happens in a home-like environment and was designed to motivate and engage users in an interaction where using the system was meaningful. We then discuss results that show significant differences among interfaces. In particular, we show that interfaces such as the smartphone interface allows nonexpert users to intuitively provide much better training examples to the robot, which is almost as good as expert users who are trained for this task and are aware of the different visual perception and machine learning issues. We also show that artifact-mediated teaching is significantly more efficient for robot learning, and equally good in terms of usability and user's experience, than teaching thanks to a gesture-based human-like interaction.**

***Index Terms*—Human–robot interaction (HRI), object visual recognition, personal robotics, robot learning, user interfaces, user study.**

Manuscript received April 20, 2012; revised August 11, 2012; accepted November 8, 2012. Date of publication December 20, 2012; date of current version April 1, 2013. This paper was recommended for publication by Associate Editor T. Asfour and Editor D. Fox upon evaluation of the reviewers' comments. This work was supported in part by the Conseil Régional d'Aquitaine and the European Research Council EXPLORERS Grant 240 007.

P. Rouanet and P.-Y. Oudeyer are with the FLOWing Epigenetic Robots and Systems Research Team, INRIA and Ensta-ParisTech, 33405 Talence, France (e-mail: pierre.rouanet@inria.fr, pierre-yves.oudeyer@inria.fr).

F. Danieau was with FLOWing Epigenetic Robots and Systems Research Team, INRIA and Ensta-ParisTech, 33405 Talence, France. He is now with Technicolor, 92443 Issy-les-Moulineaux, France, and with the VR4i Team, Institut national de recherche en informatique et en automatique, 35042 Rennes, France (e-mail: fabien.danieau@inria.fr).

D. Filliat is with the Ecole Nationale Supérieure de Techniques Avancées ParisTech, 91762 Palaiseau, France, and also with FLOWing Epigenetic Robots and Systems Team, INRIA and Ensta-ParisTech, 33405 Talence, France (e-mail: david.filliat@ensta-paristech.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TRO.2012.2228134

Fig. 1. Using a device as a mediator object between the human and the robot to control the movements of a personal robot allows nonexpert users to teach it how to recognize new visually grounded objects.

I. Introduction

A. One Challenge of Personal Robotics: Learning From Nonexpert Humans

PERSONAL robotics has been drawing an increasing amount of interest recently, both from an economic and a scientific point of view. Many indicators seem to show that the arrival of this kind of robot in our everyday homes will be one of the major events of the 21st century [1]. In particular, they are predicted to play a key role in our aging society and, especially, in applications such as domestic services, telesurveillance, or entertainment [2]. Yet, many challenges still need to be addressed before allowing personal robots to operate in our homes. They include a diverse set of questions related to perception, navigation, manipulation, learning, human–robot interaction (HRI), usability, and acceptability. In this paper, we are more particularly interested in the transverse challenge: providing the robot with the ability to adapt itself to its environment through learning by interaction with non-expert users (as illustrated In Fig. 1). This is a key feature for the development of personal robotics. Indeed, unlike industrial robotics where the environment is very structured and known in advance, personal robots will have to operate in uncontrolled, unknown, and/or changing environments. More importantly, they will have to interact with humans who may potentially have very diverse expectations and preferences. Thus, the robot should have the capacity to learn from nonexpert humans.

Fig. 2. To allow users to designate a particular object to a robot in a cluttered environment, we need to provide them with a robust and accurate pointing detection. Otherwise, it may lead to restrictive interaction and even to false learning examples.

B. Studying the Role of the Interface for Social Robot Teaching of New Visual Objects

Techniques allowing robots to learn from interaction with humans have been widely explored in the literature, including approaches such as imitation learning and learning by demonstration (e.g., [3] and [4]) or socially guided exploration (e.g., [5]). Having robots learn from humans requires both the development of machine learning algorithms (e.g., to encode and generalize new capacities) and the elaboration of intuitive and robust HRI techniques. While those two challenges are crucial, a large part of the work done in social learning focuses on the first problem [4], [6], [7]. Yet, the interaction mechanisms are known to play a key role in human teaching (e.g., [8]). Thomaz and Breazeal have shown the importance of understanding the human teacher/robotic student relationship in developing learning algorithms suited for social learning [5]. Calinon and Billard have proposed the development of learning by demonstration systems that take into account the interaction scenario [9]. Mechanisms such as joint attention have also been identified as crucial in social learning for both humans and robots teaching [10], [11]. Furthermore, the importance of the role of interfaces and interaction becomes paramount when it comes to deploying robot learning systems outside the laboratory, where they shall be used by nonexpert humans users.

In this paper, we focus on this latter issue and study the impact of human–robot interfaces, allowing nonexpert users to efficiently and intuitively teach a robot to recognize new visual objects. This is a case-study task bound to be needed in many future personal robotics applications. We present an integrated system that combines machine learning techniques, computer vision techniques, and various alternative human–robot interfaces. The goal of the whole system is to allow nonexpert humans to show new visual objects to a robot (for which it does not already have a model and, thus, cannot segment easily) and associate a name so that it can be used as a training example, allowing the robot to recognize these objects later (see–Fig. 2). A strong constraint is that the system should be efficient and usable by nonexpert users, which will provide only very few training examples per object class. It is important to notice that by "visual objects" we are not only referring to actual physical objects (e.g., a ball) but also to any region of an image having specific visual features. This very generic definition also includes more abstract objects, such as a painting, stairwell, or even an open door, which should also be recognized by a personal robot.

As we will explain, a key challenge is that nonexpert users typically have a wrong *a priori* understanding of what the robot sees or does not see, which can easily lead them to provide low-quality training examples (e.g., examples where the objects they want to show to the robot are not even on the image perceived by its camera). We argue that the design of interfaces can be key in helping nonexpert users to collect good learning examples and, thus, improve the performance of the overall learning system.

After detailing the related work in Section II, we present in Section III four alternative human–robot interfaces: three are based on the use of a mediating artifact (smartphone, wiimote, wiimote and laser) and one is based on natural human gestures (with a *Wizard-of-Oz* recognition system). These interfaces mainly vary in the kind of feedback provided to the users, permitting them to understand more or less easily what the robot is perceiving, and, thus, guide their way of providing training examples differently. As a consequence, as we will show, interfaces that provide the right kind of feedback can allow, at the same time, the human to understand what the robot is seeing at any given moment, and, vice versa, the robot can infer efficiently what the human is trying to show to it. This form of synchronization of what each other is looking at, and made possible by particular kinds of user interfaces, is an elementary form of joint attention,[1] which has been shown to be crucial in the field of developmental robotics to teach new visual concepts to a robot [11], [12]. Our study is thus located at the crossover of three important research domains: social learning in robotics [4], HRI [13], and developmental robotics [14].

We then evaluate and compare in Section IV the impact of these interfaces, in terms of learning efficiency (robot's point of view) and usability and user's experience (user's point of view), through a real world and large-scale user study (107 participants). In this study, which took place in a science museum in Bordeaux, we asked participants to teach a humanoid robot Nao[2] 12 different new visual objects in the context of a robotic game. This robotic game happened in a home-like environment and was designed to motivate and engage users in an interaction where using the integrated system was meaningful.

We chose to follow a very standard user-centered approach to design our interfaces based on mediator objects as we wanted our system to be effectively usable by nonexpert humans in plausible interactions, i.e., outside of the lab, with a personal robot. With such an approach, we first analyzed the context of

[1]It is an elementary form of joint attention in the sense that both agents can infer what the other is looking at and/or perceiving, without an explicit cognitive model of attention; see [11].

[2]The Nao robot represents, in our opinion, the current personal affordable robots well, with a humanoid appearance. Furthermore, we choose to use it as an autonomous robot, i.e., with only onboard sensors, and not to enhance its capacities with external devices, such as cameras fixed on the ceiling such as those used in smart environments or ubiquitous robotics approaches [15]. We argue that the complexity of this kind of installation could prevent their use in everyday homes in the near future. Second, it is important to note that this kind of system, while improving the perceptual capacities of the robot, will not fundamentally change the attention problem that we are trying to tackle here. For instance, a pointing gesture will remain ambiguous in a cluttered environment, even with the use of fixed cameras.

use, then conceived the interface, and finally, we evaluated it. This cycle was repeated until the specified requirements were matched. In this paper, we are presenting the last complete iteration of our development. While some earlier versions of subparts of our integrated system have already been presented in [16]– [19], they have been modified and improved since then and are here presented as an integrated system for the first time.

We then discuss results that show significant differences among interfaces. In particular, we show that interfaces, such as the smartphone interface, allow nonexpert users to intuitively provide much better training examples to the robot, almost as good as expert users who are trained for this task and aware of the different visual perception and machine learning issues. We also show that artifact-mediated teaching is significantly more efficient for robot learning, while better in terms of usability than teaching using gesture-based human-like interaction.

Finally, a discussion of the main results and of our design choices is presented in Section VI.

II. Related Work

The classification and recognition of new visual objects have been studied intensely from visual perception and machine learning perspectives. Many approaches, such as the *bags of visual words* we are using in this paper, have been recently developed [20]– [22]. Those learning systems are highly efficient when trained with a large database of good labeled examples (see, for instance, PASCAL VOC [23]). Yet, to solve this problem in a real HRI scenario, i.e., outside of the laboratory, with nonexpert users in a realistic use scenario, one needs to tackle a crucial issue not addressed in the machine learning and computer vision literature: how to collect good training examples through relatively few but intuitive interactions with nonexpert users? And, how to collect examples by using current social robots that typically have limited sensors and, in particular, a strongly constrained visual apparatus? Those questions are addressed in this paper.

The questions of drawing a robot's attention, which is pointing toward objects, and realizing various forms of joint attention to teach the name of new objects have also been widely studied. For instance, they are closely related to research done in robot language acquisition and in particular the construction of visually grounded lexicons [24]–[26]. Yet, in this literature, most authors are focusing on the perception and machine learning questions. In particular, as they try to model human language acquisition, they choose to directly transpose the human-like interactions to HRIs to allow humans to show new associations between visual objects and their names. For instance, Kaplan developed a complete social framework based on human-like interactions, such as pointing gestures and speech recognition, to allow users to teach words associated with objects to an AIBO robot [27]. Scasselati used pointing gestures and gaze tracking to draw a robot's attention [28]. In this study, he used a fixed upper torso and, thus, constrained the interaction. Pointing gestures have also been used to guide a robot companion [29], [30].

Unfortunately, existing associated techniques for gesture, gaze and speech recognition, and interpretation 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 these modes of interaction (low quality and noisy sensor, small angle of view, small height, etc.). Thus, these *a priori* intuitive systems have to be used by expert users in the sense that they have to understand the limitations of the robot in order to behave according to a very restrictive protocol which will allow the interaction to work. One way to circumvent this problem is to have a very controlled setup. For instance, Roy presented a framework that allows a robotic system to acquire visually grounded words [31]. Here, users have to place objects in front of the robot and then describe them. We argue that this kind of experiment cannot be directly transposed into a real-world application in personal and social robotics with nonexpert users.

Yet, as personal robotics is predicted to become commonplace in our home environments in the 21st century, it is really important that even nonexpert users can robustly designate objects to their social robot in an uncontrolled environment. We should provide interfaces which are intuitive in order to avoid misunderstanding or frustration during interaction but also to help users collect good learning examples. Indeed, in a cluttered environment, nonrobust pointing may lead to the designation of the wrong object and, thus, completely incorrect learning examples which will decrease the performance of the whole learning system. In their work, Kaplan and Steels identified the lack of robustness in the interface as a major limitation of their system and showed that the lack of robustness of the interface often leads to a number of bad learning examples [32].

Another widely used way to tackle this pointing and joint attention problem is to allow users to directly wave objects in front of the camera of the robot [33], [34]. Thus, we can ask the robot to always focus its attention on the moving objects. Furthermore, it also allows the separation of the object from the background by subtraction of the motionless part of the scene. However, with this technique, users can only show to the robot small and light objects that can be easily carried as they will have to be waved in front of the robot. Thus, we cannot show objects such as a table, a plug, or a painting on a wall. Moreover, for the elderly or the disabled, waving objects could be really tiring or even impossible.

We argue that one way to help achieve some of the abilities described previously intuitively and robustly without facing the problems encountered when waving objects is to develop simple artifacts that will serve as mediators between the human and the robot to enable intuitive communication. Interfaces that are based on mediator objects have already widely been used in the domain of HRI and, especially, to draw a robot's attention toward an object. For instance, Kemp *et al*. used a laser pointer to easily and robustly designate objects to a robot in order to ask it to fetch them [35]. Here, they used the laser pointer as a point-and-click interface. They showed that inexperienced participants managed to correctly designate objects to a robot. Furthermore, 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 that will be grasped by a wheelchair-mounted robotic arm [36]. In

their study, the user can directly monitor the object selection on the screen of the device. As in our system, they can both draw the robot's attention toward objects and, therefore, realize joint attention between the human and the robot. However, their robot is able to automatically grasp the object from a detected 3-D spot in a framework that requires an image segmentation algorithm and/or *a priori* object knowledge. If objects are not known beforehand, these are still difficult problems.

Other mediator object-based interfaces have been developed recently. For instance, Fong *et al*. used a personal digital assistant for remote driving [37], and Kaymaz *et al*. used it to teleoperate a mobile robot [38]. Sakamoto *et al*. showed how they can control a house cleaning robot through sketches on a Tablet PC [39]. Ishii *et al*. proposed a laser pointer-based interface where users can draw stroke gestures using the laser to specify various commands such as path definition or object selection with lasso gestures [15]. However, in their work, they used calibrated ceiling-mounted cameras and vision-based ID tags to circumvent object recognition issues. Yet, to our knowledge, nobody has used this kind of interface for interactions that involve robot teaching, such as teaching new words for new visual objects.

III. Outline of the System

As explained previously, we present here an integrated system to allow nonexpert users to teach a personal robot how to recognize new visually grounded objects in real-world conditions. In particular, this means that our system should allow a user to draw the robot's attention toward an object present in its surrounding and then collect a learning example of it. The robot could thus recognize and search for an already taught object later on. In this version of the system, labels are automatically associated with images. We will describe in Section VI a more advanced version of our system, which allows users to associate new acoustic words with the visual objects.

This system has to deal with visual perception, machine learning, and interaction challenges. The visual perception and machine learning parts of our system are based on a version of the advanced *bags of visual words* technique [21]. These computer vision and machine learning algorithms have been chosen because, to us, they represent robust and standard tools often used as a baseline to compare with more recent techniques. Furthermore, we are here focusing on the four different interfaces notably developed to tackle the pointing and attention challenges. Three interfaces are based on mediator objects, while the last one is based on arm and hand gestures with Wizard-of-Oz recognition.

Our system was embedded in the Nao robot that is designed by the company Aldebaran Robotics.[3] The robot was only used here to collect the learning examples (i.e., take the pictures) and store them. The actual learning was performed offline on a computer. We have already explained why we chose this particular robot and used it as an autonomous robot. The implication of this choice will be discussed later.

[3]http://www.aldebaran-robotics.com/

A. Visual Perception

We adopted the popular *bags of visual words* approach [20] to process images in our system. This method was developed for image categorization and object recognition and relies on a representation of images as a set of unordered elementary visual features (the words) taken from a dictionary (or code book). The term "bag of words" refers to text document classification techniques that inspired this approach where documents are considered to be an unordered sets of words. In its basic implementation that we use here, a classifier that predicts the object identity is based on the occurrence frequencies of the visual words in an image, thus, ignoring any global image structure. There are several extensions that introduce some global geometry in order to improve performance (e.g., [40]), but these extensions were not necessary to implement in order to demonstrate the interest of the interfaces, which is the subject of this paper. Several applications also exist for robotics, notably for navigation (e.g., [41] and [42]).

The words used in image processing are based on automatically detected local image features. The feature detectors used are usually invariant to image rotation, scale, and partially to affine deformation to be able to recognize objects under varying point of view. Among the many existing feature detectors, we chose speeded up robust features [43] for its performance and reasonable processing cost. For each detected feature, a descriptor is computed that encode the local image appearance. A dictionary is created by clustering a large set of feature descriptor extracted from images representative of the environment. In our implementation, we use a hierarchical k-means algorithm to create a tree-structured dictionary that enable fast word lookup [44]. The size of the dictionary was set to 2^{12} in our experiments.

This model has interesting characteristics for our application: The use of feature sets makes it robust to partial object occlusions and the feature space quantization brings robustness to image noise which is linked to object position, camera noise, or varying illumination.

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 involves evaluating the likelihood of each object given a new image.

More specifically, we use a voting method based on visual words occurrences for each object. The recorded statistics during learning (according to the learning method described later) are the number of occurrences O_{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 $O_{wo} \neq 0$. The vote is performed using the *term frequency–inverted document frequency (tf–idf)* weighting [20]

in order to penalize the more common visual words. The recognized object is the one with the best vote.

Estimating the statistics O_{wo} requires the labeling of examples with their associated object name. The quality of object recognition is obviously strongly influenced by the number and quality of training images [45]. In computer vision, creating good image datasets is therefore an important aspect, to which a large amount of time is devoted, but, this time, is not available when interactive learning takes place with a robot for new objects as performed in our study. Moreover, precisely selecting relevant training images is also not always possible, depending on the interface used to control the robot. As will be described later, we will use two methods for labeling based on the information given by the user: labeling the whole image or labeling only an image area (given by the user) that represents the entire object. Then, we will show the influence of these methods on final object recognition.

C. Human–Robot Interaction

In this section, we present the different interfaces developed. They were chosen to span the variety of mediator interfaces that one can imagine but also to explore the different kinds of feedback of what the robot is perceiving that can be provided to the users. Three of the interfaces are based on mediator objects such as the iPhone, the Wiimote, or the laser pointer. We chose rather well-known and simple devices; therefore, users can quickly learn how to use them. The fourth interface was added in order to compare the mediator-based interfaces with a human-like interaction which, as we will demonstrate, reveals itself to be less usable and less efficient than the mediator-based interfaces.

In order to be compared fairly, each of these four interfaces has to provide the users with the exact same following abilities:

1) driving the robot;
2) drawing its attention toward a direction or a specific object;
3) defining the object area inside the image (only the iPhone and Wiimote laser interfaces provide this ability; for the two other interfaces, the whole image was taken into account in the evaluation).

The mediator objects were not used to trigger the capture of a new learning example. Instead, when users think that the robot sees the object they want to teach, they had to directly touch its head. We chose to force this physical interaction with the robot in order to increase the feeling of collaboration. Yet, the different mediator objects could easily be adapted to directly trigger the capture.

It is important to notice that all the interfaces were based on the exact same sensorimotor capacities and functionalities of the Nao robot. As argued previously, the Nao sensorimotor apparatus represents well the present form of existing social robots to us. We also voluntarily choose not to enhance its capacities by using a ceiling or a wide range camera, although it may have improved the usability of our interfaces. Indeed, as discussed in detail in Section VI, such an improvement would not have solved the fundamental attention problems that we are trying to tackle here.

In the next sections, we will describe each interface in detail and emphasize their specificities. We chose to focus on three interfaces, which seemed the most interesting to us, but other mediator objects or interaction metaphors could have been imagined in this context (e.g., using a pico projector to display the field of view of the robot could have helped knowing what the robot could see).

D. iPhone Interface

The first interface is based on an Apple iPhone used as a mediator object between the human and the robot.[4] We chose this device because it allows the display of information on the screen to the user as well as interaction through intuitive and standard gestures. In addition, the multitouch capacity provides numerous possibilities. Due to the large success of the iPhone, we can take advantage of a familiar interface, allowing ease of use.

In this system, the screen of the handheld device displays a continuous video stream of the robot's camera. It accurately shows what the robot is looking at, which can be monitored by the user who resolves the ambiguity of what the robot is really seeing (see Fig. 4). However, the user's attention is split into direct and indirect monitoring of the robot which may lead to the increase of the user's cognitive workload. Finally, having visual feedback seems to entertain the user, while the robot is moving, as shown in pilot studies [16].

When the human wants to draw the robot's attention toward an object, which is not in its field of view, the user can sketch on the screen to make it move to an appropriate position: vertical strokes for forward/backward movements and horizontal strokes for right/left turns. Elementary heuristics are used to recognize these straight touch gestures. The moves of the robot are continuous until the user retouches the screen in order to stop it. Pointing on a particular point on the screen makes the robot look at the corresponding spot (see Fig. 5). This is a very convenient way of drawing the robot's attention toward a specific object.

As explained previously, when the user wants to show an object in order to teach a name for it, he can first make sure that the object is in the field of view of the robot by monitoring whether the object is displayed on the screen or not. Once he is certain that the robot sees the object, he touches the head of the robot to ask it to take a picture. Then, the system asks the user to sketch a circle around this object directly on the touch screen (as shown in Fig. 6). 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 [46]. Hachet *et al.* used 2-D circle inputs for easy 3-D camera positioning [47]. As for the straight strokes, heuristics are used here to recognize circular touch gestures based on the shape of the stroke and the distance between the first and the last point of the gesture. Circling is crucial to the robot since it provides a rough visual segmentation of the object, which is otherwise a very hard task

[4] http://youtu.be/vrMsaIj2SDM

Fig. 3. To make the robot collect a new learning example, users have to first draw the robot's attention toward the object they want to teach through simple gestures. Once the robot sees the object, they touch the head of the robot to trigger the capture. Then, they directly encircle the area of the image that represents the object on the screen. The selected area is then used as the new learning example. (a) Draw the attention toward an object. (b) Trigger the capture. (c) Encircle the area of the object. (d) New learning example.

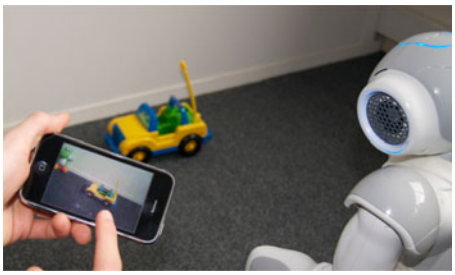

Fig. 4. Video stream of the camera of the robot on the screen. This allows accurate monitoring of what the robot is seeing.

Fig. 5. Drawing attention toward 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, the user can just tap on the screen (right).

Fig. 6. Once the user asks the robot to take a picture of a new object, he can directly encircle it, thus providing a useful rough object segmentation.

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). The complete sequence of actions needed to provide a new learning example is summarized in Fig. 3.

E. Wiimote Interface

The second interface is based on a Wiimote device (see Fig. 7).[5] The users can press one of the buttons of the arrows to move the robot. We use the very common flying vehicle

Fig. 7. Users can move the robot by using the directional cross or directly orienting the Wiimote to aim its head. However, the lack of feedback makes it very difficult to estimate whether the robot really sees the object the user wants to teach.

metaphor: If we want to make the robot move forward, we keep the up arrow pressed, and as soon as the button is released, the robot will stop. This technique permits easy driving of the robot or the ability to draw its attention toward a general direction. To aim the head of the robot, users have to orient the Wiimote, i.e., we directly map the values of the accelerometers to the pan/tilt values of the robot's head. Thus, users can always focus their attention on the robot.

However, this interface does not provide any feedback about what the robot is perceiving. In particular, users cannot be sure whether the robot sees an object or not. Therefore, they have to "guess" what the robot really sees, which can be a very difficult task, as illustrated in the experiments presented in the following.

F. Wiimote and Laser Interface

In this third interface, the Wiimote is also used to drive the robot. However, as shown in Fig. 8, a laser pointer is combined with the Wiimote and used to draw the robot's attention.[6]

The robot is automatically tracking the laser spot and aims its head in order to keep the spot near the center of its sight. We chose a laser pointer as the interaction device as this method to draw someone's attention is quite common in our everyday life, in oral presentations for instance, and therefore is an intuitive interaction for users. Here, users can draw the robot's attention toward a direction by smoothly aiming its head toward the right direction or they can point to a particular object directly once

[5]http://youtu.be/vrMsaIj2SDM

[6]http://youtu.be/vrMsaIj2SDM

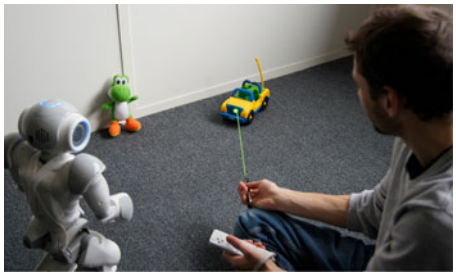

Fig. 8. Users can drive the robot with a Wiimote and draw its attention toward an object by pointing at it with a laser pointer as the robot is automatically tracking the laser spot.

inside the robot field of view by designating it with the laser pointer.

We automatically detect the laser spot in the images received from the camera of the robot. We used a very bright laser pointer with a significant spot size. We chose a green pointer because "laser" green is probably a color that is less present in the everyday environment and, therefore, much more salient.

Unlike Kemp *et al*. who used an omnidirectional camera [35], or Ishii *et al*. who used ceiling-mounted cameras [15], in our study, the field of view of the robot is very limited. Therefore, drawing the robot's attention requires that the user correctly estimates the field of view of the robot. This can be a difficult task as nonexpert humans are often prone to assume that the robot has a field of view which corresponds to a human one, but this is not the case most of the time.

We provide different feedback to help users better understand when the robot is detecting the laser spot and, thus, correctly estimate the field of view of the robot. First, as the robot is tracking the laser, users can monitor the movements of its head so that they have visual feedback. Second, we also provided haptic feedback by vibrating the Wiimote each time the laser spot was detected by the robot. With the combination of these two feedbacks, users know whether the robot was detecting the laser spot or not and can make sure that the laser pointer and, thus, nearby objects are in the field of view of the robot.

The feedback for this interface is more restricted than the complete feedback that is provided by the iPhone interface where users can directly monitor what the robot sees. Furthermore, users cannot be sure that the robot sees the object they are designating entirely. They can only be sure that the robot is detecting the visible spot on a part of the object. For instance, when it is not possible to center the robot's head on the laser spot due to the robot's physical limit, a part of the object may be outside of its field of view.

The laser also provides visual feedback to the user. Indeed, the laser spot allows users to monitor what they are really pointing at, and thus, they can adjust their pointing if needed. This is particularly important in a cluttered environment where small deviations may lead to pointing to the wrong object, leading to an incorrect learning example.

Once users manage to draw the robot's attention toward an object and trigger the capture of a learning example by touching the head of the robot, they can then encircle the object directly with the laser pointer (see Fig. 9 for the whole sequence). To record the encircling gestures done by the user, we store the detected points during these movements. Yet, as the frame rate of the camera is low and as the speed of the movement may really vary from one person to another, encircling once was not always enough to compute a reliable stroke. Therefore, we asked participants to encircle the objects many times. All the detected points are recorded without keeping any structure information and stored in a point cloud. It is then fitted on an ellipsis, as the shape of encircling gestures tends to be elliptic. Finally, the robot indicates through a head movement whether it has detected enough points to compute a reliable ellipsis. The ellipsis is computed as follows [48]:

Based on the implicit equation of an ellipsis, we can obtain the following system:

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0 \text{ with } A \neq 1$$
$$\Rightarrow Bxy + Cy^2 + Dx + Ey + F = -x^2.$$

Written in a matrix form

$\alpha X = \beta$ with

$$\alpha = \begin{pmatrix} x_1 * y_1 & y_1^2 & x_1 & y_1 & 1 \\ \dots & \dots & \dots & \dots & \dots \\ x_n * y_n & y_n^2 & x_n & y_n & 1 \end{pmatrix} \quad \beta = \begin{pmatrix} -x_1^2 \\ \dots \\ -x_n^2 \end{pmatrix}.$$

As this system is overdetermined, we try to find the X which best fits the equation in the sense of the quadratic minimization problem (least squares)

$$\arg\min_X = \left\| \beta - \alpha X \right\|^2$$

which is equivalent to solving the equation

$$\hat{X} = (\alpha^T \alpha)^{-1} \alpha^T \beta.$$

Once the ellipsis has been computed, we can use it to delimit the boundary of the object and, thus, roughly segment the image. In opposition to the iPhone interface where the encircling is done on the 2-D image, here users encircle directly on the 3-D environment. This could lead to projection issues, especially, when the background is not planar. As we can see in the examples in Fig. 10, the projected stroke could sometimes "cut" the object and, thus, decrease the quality of learning examples.

G. Gesture-Based Interface With Wizard-of-Oz Framework

In this last interface, users can guide the robot by making hand or arm gestures (see Fig. 11).[7] As we wanted to keep this interface as intuitive as possible, we did not restrict the kinds of gestures that users can make. However, as gesture recognition is still a complex task, we used a Wizard-of-Oz (WOZ) framework where a human was controlling the robot according to different gestures the magician recognized. We can thus ensure that the recognition of the different gestures was not a limitation of this interface. As stated previously, we did not want to enhance the robot's capacities with a ceiling or wide-angle

[7] http://youtu.be/l5GOCqXdgQg

Fig. 9. With this interface, users can draw the robot's attention with a laser pointer toward an object. The laser spot is automatically tracked by the robot. They can ensure that the robot detects the spot, thanks to the haptic feedback on the Wiimote. Then, they can touch the head of the robot to trigger the capture of a new learning example. Finally, they encircle the object with the laser pointer to delimit its area, which will be defined as the new learning example. (a) Draw the attention toward an object. (b) Trigger the capture. (c) Encircle the area of the object. (d) New learning example.

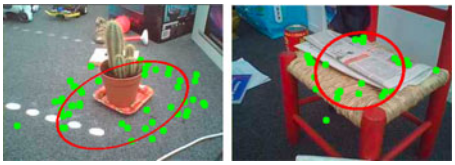

Fig. 10. Encircling with a laser pointer raises difficulties mostly due to the projection of the laser spot in the plane of the camera.

camera as we wanted to study a human-like interaction between nonexpert users and what we think represent a typical actual personal robot well. Thus, all the four interfaces are based on the same robot sensorimotor capacities which allows a comparison of the different interfaces on a fairer basis. Furthermore, as explained in the discussion, even if using external sensors would probably improve the usability of this interface, it would not fundamentally change pointing or attention issues.

As the Wizard was only seeing interaction through the robot's eyes and therefore through a very limited visual apparatus, most of the gestures made by the users were outside of the robot's field of view. As a consequence, and as we will show in the experiments in the following, even such a well-known interface with human-level intelligence may in fact lead to fragile interactions due to the differences between the robot's sensorimotor apparatus and the human's ones.

Obviously, this interface embeds a strong difference with the others. Indeed, the Wizard is a human who already has knowledge and strong biases about what may constitute an object shown to him, i.e., the object segmentation problem is here automatically solved by the human. Thus, when the object was pointed at, the wizard naturally centered the sight of the robot on it. Yet, on the other hand, in spite of this advantage when compared with other interfaces that are coupled with an autonomous robot, this interface does not perform so well, as we will demonstrate.

Although this interface embeds some strong differences with the other, it still appears very interesting to us, as it first allows the investigation of a human-like interaction with an autonomous personal robot with a limited visual apparatus if we assume a human-level recognition and interpretation of gestures. Second, it also permits the comparison of our interfaces based on mediator objects with a human-like interaction and showed that because of the particular visual apparatus of the robot, this interaction may lead to a more restrictive and, thus, less satisfying interaction for users.

IV. Experimental Evaluation With Nonexpert Users

As previously explained, we want here to study the impact of interfaces on robot learning of new visual objects through nonexpert user teaching and in a real-world home-like environment. This impact is evaluated along two main dimensions.

1) *Learning efficiency:* We test the quality of training examples collected by the robot through human guidance. This quality is evaluated both qualitatively (see later) and quantitatively through a measure of accuracy in classification in generalization (i.e., on images not in the collected dataset). We also want to study more specifically how encircling can impact the performance of the learning system.
2) *Usability and user's experience:* We study in the following how intuitive, effortless, and entertaining our interfaces are for nonexpert users.

We argue that potential future users of social robots will not necessary be expert users, in the sense that they should be able to interact with their social robot without any prior specific training. Thus, it is crucial that our study is not a laboratory study but a real-world study. In particular, we want to create a plausible context of interaction and have representative nonexpert participants, in order to preserve the ecological validity of our results and avoid the classical pitfalls of the evaluation in the HRI domain [49], [50].

However, as those users have probably never interacted with a social robot before, asking them to show and teach objects to a robot is still an unusual and artificial task, as shown by pilot studies [16], [17], [51]. Therefore, we need to embed this task in a scenario in order to justify it. In particular, we need to encourage the users to collect high-quality learning examples. Moreover, we want a scenario that can entertain and maintain the user's motivation during the whole experiment. Finally, we want to conduct a large-scale study, and therefore, we need to design a formal and reproducible experiment.

A. Robotic Game Experiment

We argue that one solution to tackle the aforementioned issues was to design our user study as a robotic game. Games are well known to be a powerful way of captivating and engaging users through their storyline. For instance, serious games have been widely used for education or training, allowing learners to experience situations that are impossible or hard to reproduce in the real world [52]. We think that, in the same way that

Fig. 11. In this mode of interaction, the robot is guided by the hand and arm gestures made by the user. In order to have a robust recognition, we used a WOZ framework, where the wizard was only seeing the interaction through the robot's viewpoint.

Fig. 12. For the experiment, we used 12 textured objects directly related to football: beer, ball, gloves, coke, a poster of Zidane, a jersey of Beckham, a poster of a stadium, a jersey of the Bordeaux team, shoes, a gamepad, a video game, and magazines. Each participant had to teach four randomly chosen objects to the robot to help it better understand football.

video games have managed to make novice users solve complex and unusual tasks by using mechanisms, such as tutorials or briefings, we could design a robotic game experiment that helps users to better understand and remember all the steps needed to teaching visual objects to a robot. The scenario of the game also permits us to justify this artificial task. Finally, presenting the experiment as a game allows us to attract a wide and varying panel of participants. Participants would feel more comfortable participating in a game than in a scientific experiment.

Thus, we created a game scenario to try to match all of the above. The users were told the following story: *A robot, which has come from another planet, has been sent to Earth in order to better understand what seems to be a popular human habit: "playing football." Indeed, from their remote home, the robots have just picked up partial information about this practice, and therefore, they want to investigate further. Therefore, one robot was sent to the living room of a football fan to gather more clues. As the robot was damaged during its journey, it could no longer fulfill its mission alone. Therefore, you will need to help it!* The user was asked to help the robot to collect clues (i.e., collect learning examples of four different objects related to football). Every time the robot collected a new example, a false and funny interpretation of what the object can be used for was given by the robot.

B. Experimental Setup

1) Game Environment: We recreated a typical 10 m^2 living room that is located next to the café of a science museum in Bordeaux, France. We arranged furniture, such as tables or chairs, and many other various everyday objects (newspaper, plush toys, posters, etc.) in order to make it look inhabited. Among these various objects, 12 were directly related to football (see Fig. 12). Those objects were the possible clues the robot needed to collect. They were chosen because they fit well within the scenario and because they were textured and big enough so that they could be robustly recognized by classical visual recognition algorithms (if provided with good quality learning examples). Other usual objects were added to the scene to make the environment cluttered (see Fig. 13).

The design of the game environment had three main purposes.

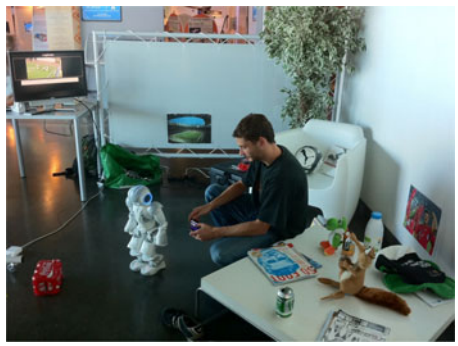

Fig. 13. Real-world environment designed to reproduce a typical living room. Many objects were added in the scene in order to make the environment cluttered.

1) Reproduce a daily life area to provide participants with a stressless environment and to reduce the feeling of being evaluated.
2) Conduct the experiment in a realistic environment; therefore, users have to navigate the robot through a cluttered area and to collect real-world learning examples (lighting conditions, complex background, etc.).
3) Immerse users in the scenario.

The global structure of the room remains unchanged during the whole experiment in order to get a constant test environment. Nevertheless, the small objects were randomly arranged every five experiments. Indeed, in a real home, while big objects such as furniture will not move, most of the small objects will often be moved and, thus, must be recognized in spite of their background.

2) Robot: As stated previously, we used the Nao for our experiment. To make it more lively, we developed some basic behaviors, such as yawning or scratching its head, if the robot was idled for a long time. We also used different colors for its eyes to express simple emotions or to provide feedback to the users (see Fig. 14). Moreover, we added organic sounds to express the robot's mood.

3) Game Interface: As explained previously, the design of our robotic game was inspired from a classic video game. We used a large screen as the game interface to display information to users such as a cutscene video explaining the story.

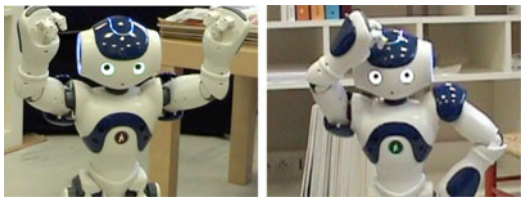

Fig. 14. Behaviors, such as "happy" (on the left) or scratching its head (on the right) were designed to make the robot look more lively and help the users better understand its behavior.

Fig. 15. Story of the game was told through a video displayed on our game interface. This display was also used to provide users with step-by-step instructions of the tutorial.

This interface was also used to recreate a tutorial where participants learn one ability at a time: walking straight, turning, aiming the head of the robot, and collecting a learning example. For each step, a short video explained how to realize the task with the interface they were using. After the video, the user was asked to effectively complete the task. Once they succeed, they could move on to the next stage. These videos were also a way to make users know the robot better. The final objective of the tutorial was to first collect a learning example which, in fact, was a picture of the user's face. The whole tutorial lasted about 5 min on average. After the tutorial, the real mission was explained to the participants via another video similar to the one in Fig. 15. Thus, the game interface allowed us to present the whole experiment (both tutorial and mission parts) in one single game.

Furthermore, it also allowed us to conduct each test in the same way. Indeed, all participants received the exact same information and instructions through the game interface.

C. Experimental Protocol

The experiment took place from June to November 2010, and 107 persons participated in it. Most of them (74) were recruited at Cap Sciences,[8] which is a science museum in Bordeaux, most of which were visitors. We expected to find, in general, nonexpert participants within the public, although it might have introduced a bias as science museum visitors are probably more receptive to technology. The others (33) were recruited on the campus of Bordeaux University of Technology. We expected to find here participants with a significant technological background and a knowledge of classical interfaces but without any particular robotic knowledge.

Seventy-seven participants were male and 30 were female. The participants were aged between 10 and 76 (M = 26.3, STD = 14.8). Among the 107 participants: 32 used the iPhone interface, 27 the Wiimote interface, 33 the Wiimote-laser interface, and 15 the gestures interface.

Each participant was asked to follow the following protocol, which was generated from the result of several pilot studies.

1) Fill in a consent form.
2) Fill in a pre-questionnaire.
3) Experimentation (robotic game).
   a) Tutorial.
      i) Wake up the robot by touching its head.
      ii) Make it move forward.
      iii) Make it turn left and right.
      iv) Turn its head left, right, up, and down.
      v) Make it watch your face (or a ball for the laser interface).
      vi) Enable the photo mode by touching its head.
   b) Mission.
      i) Draw the robot's attention toward one randomly chosen object among the 12 other possible objects.
      ii) Take a picture of it.

   The steps from i) to ii) were repeated four times.
4) Fill in a post questionnaire.

The whole experiment (including the questionnaires) lasted from 20 to 30 min per participant.

D. Measures

During the experiments, we collected the pictures taken by the robot and analyzed them, as described in the following. Due to the nature of the game interface, the images were automatically labeled. Indeed, the participants were asked to show a particular object that was indicated by the game interface. We also measured the time needed to complete the game as well as the intermediate times, i.e., each time a picture was taken.

On top of these measures, we also conducted two questionnaire-based surveys inspired by the classical guidelines found in the HRI literature [49], [53]. Before the experiment, we administered a demographic survey and a pretask questionnaire concerning the participant's technological profile (computer, video games, and robotic experience) and their attitude toward robotics. After the game, we conducted a posttask survey with the following assertions to which agreement had to be evaluated on a five points Likert scale.

1) Usability and user's experience.
   a) It was easy to learn how to use this interface.
   b) It was easy to move the robot.
   c) It was easy to make the robot look at an object.
   d) It was easy to interact with a robot.
   e) The robot was slow to react.
   f) Overall, it was pleasant to use this interface.
2) Robotic game.
   a) Completing the game was easy.
   b) The game was entertaining.
   c) I felt like cooperating with the robot.
   d) I picture myself playing other robotic games in the future.

[8] http://www.cap-sciences.net/

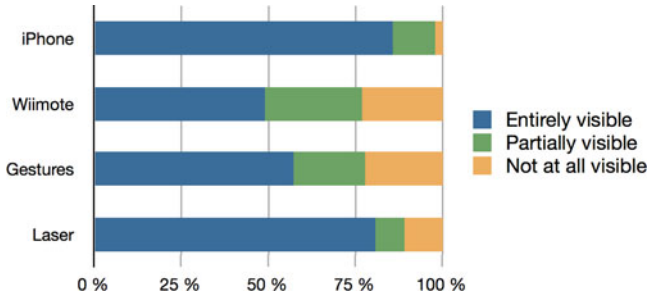

Fig. 16. Partition of the collected images into three categories: The object is 1) entirely, 2) partially, or 3) not at all visible on the images. We can see that without any feedback (Wiimote or Gesture interfaces), the object was entirely visible in only 50% of the examples. Providing feedback significantly improves this result (80% for the laser and more than 85% for the iPhone).

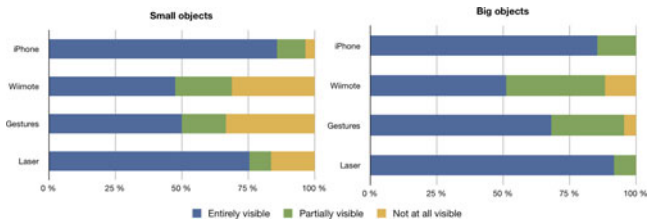

Fig. 17. Similar charts to Fig. 16, but here, the collected images were split into two subsets: small and big objects. We can see that the difference between the interfaces is even more accentuated for small objects: With the Wiimote and Gesture interfaces,' participants failed to provide correct learning examples in about one third of cases.

V. Results

A. Qualitative Analysis of the Images

We first manually sorted the collected pictures, i.e., the visual training examples corresponding to a new object, into three categories: 1) images where the object was entirely seen; 2) images where the object was only partially present; and 3) images where the object was not present at all. The objects were defined as "partially visible" as soon as a part was missing. We here considered the object corresponding to the label obtained, thanks to the game interface. Fig. 16 depicts these results. We performed a one-way analysis of variance (ANOVA) where the independent variable was the interface used and the dependent variable was the number of object corresponding to the "entirely visible" condition. We found a statistical difference between the four interfaces ($F_{3,103} = 13.7, p < 0.001$). In particular, we can observe that without providing any feedback about what the robot sees to the users (the Wiimote and gestures conditions), the object is entirely visible in only 50% of the images. The Tukey posthoc test showed that providing feedback significantly improves this result (80% for the laser and 85% for the iPhone). Furthermore, we can discern that the iPhone interface and in particular its video feedback prevents users from collecting incorrect learning examples (i.e., where the object is not present) in most cases (only 2%).

We also split the images into two subsets:

1) big objects: the two posters, the two jerseys, and the ball;
2) small objects: the other objects.

As we can see in Fig. 17, the differences between the interfaces are more accentuated for small objects. In particular, we can see that the lack of feedback led to about one-third of examples being incorrect. While the laser feedback improves this result (an error rate of only 20%), only the iPhone interface seems to really prevent users from providing incorrect learning examples regarding small objects. Finally, we can also observe that users managed to provide rather good examples of the big objects across all the interfaces. Yet, while the objects were almost always entirely visible under conditions where they used the iPhone and Laser interfaces (more than 85% of the case), they were only partially visible in about one third of the examples under the Wiimote and Gesture conditions.

B. Quantitative Analysis of the Images

We also used the collected images as input training for our learning system in order to have an effective measure of the quality of the learning examples and their impact on the overall performance. As explained previously, our learning system is based on a *bags of visual words* approach. This technique is based on a dictionary that is used to categorize the features extracted from the images. For our tests, we built a dictionary by recording a 5-min sequence using the camera of the Nao (about 1000 images), while it was navigating in the living room of our lab. We ensured that none of the furniture or the objects used during the experiments were present during the recording in order to have a dictionary that was not specific to our experimental setup.

We used the following protocol for all the tests.

1) We randomly chose N input images per object collected by users who used a specific interface. Thus, we mixed images taken by several users. As collecting one example of five objects already took about 20–30 min per participant, we could not ask them to collect few examples of each of the 12 objects. This certainly introduced a bias that we tried to counterbalance by randomly selecting images and repeating our tests many times. As shown in the results below, the variability between users (the standard deviation) is rather low. In particular, in most cases, the variability among interfaces is larger than the differences among the users of an interface.
2) We trained our learning system with these input images.
3) We tested our learning on the test database (see later).
4) The test was repeated 50 times by choosing a different set of N training images each time in order to counterbalance the randomization effect.
5) The final results are the mean recognition rate and variances of each test.

The test database was built by an expert user who collected ten examples of each of the objects through the Wizard interface. The images were taken in the same experimental setup as the actual experiment. These examples represented the "best" examples that we could expect as an input. The database was then split in half: the first part was used as a "gold" training input, while the other half was used as the test data. These "gold" examples were to be used in a similar way as the examples collected with one interface. They were to provide us with an interesting baseline to determine which recognition rate our learning system can achieve with such optimal examples.

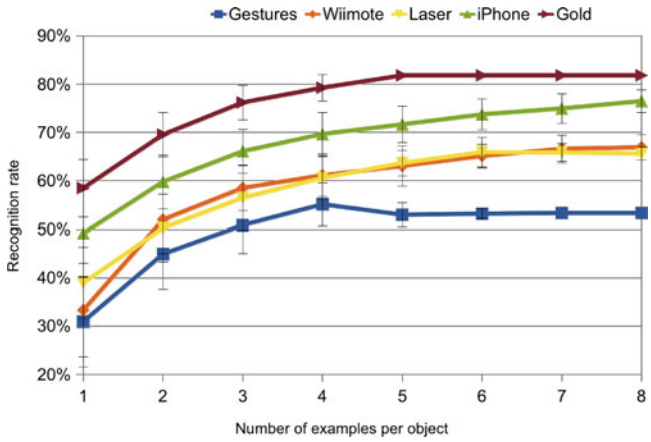

Fig. 18. *Recognition rate for all the 12 objects:* Impact of the interface on the quality of the learning examples and, therefore, on the generalization performance of the overall learning system. In particular, we can see that the iPhone interface allows users to collect significantly higher quality learning examples than the other interfaces. Furthermore, it allows even nonexpert users to provide the learning system with examples of a quality close to the "gold" examples provided by an expert user.

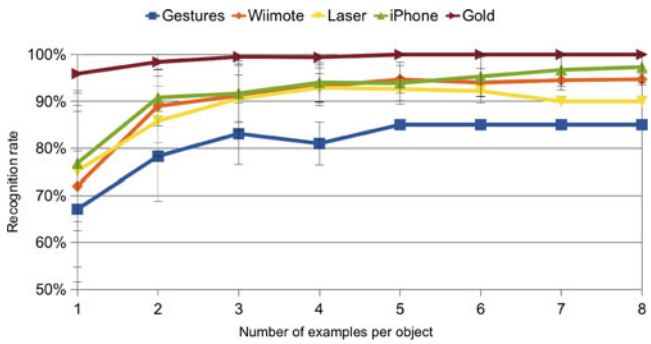

Fig. 19. *Recognition rate for the five big objects*: We can see that all the mediator interfaces allow users to collect equally good learning examples. Therefore, for the big objects, the interface does not seem to have a strong impact on the recognition rate.

As shown in Fig. 18, we notice first that the examples collected with the iPhone interface led to a significantly higher recognition rate than the other three interfaces. In particular, we notice that only three examples collected with the iPhone give as good results as eight examples of any other interface. Furthermore, the iPhone interface seems to allow nonexpert users to achieve results that are close to the results achieved with the gold training after eight examples. We can also see that even with very few good learning examples (such as three or four iPhone examples), we can achieve a rather high recognition rate of 12 different objects (about 70% correct recognition). Then, we can see that the lowest score was obtained with the Gesture interface. This result can probably be explained by the lack of usability of this interface (see details in the next section).

As in the previous section, we also separated the 12 objects into two categories: big or small. As can be seen in Fig. 19, the recognition rate for the big objects is very high (about 90%) for all the mediator interfaces. Furthermore, we can see that no significant difference was found between these interfaces. On the other hand, we can see in Fig. 20 that for the small objects,

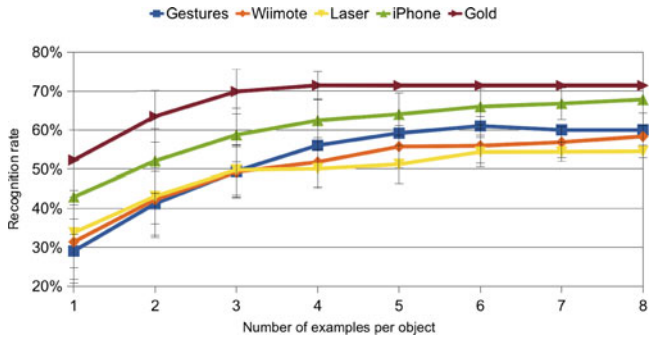

Fig. 20. *Recognition rate for the seven small objects*: We can see that the iPhone interface allows users to provide higher quality learning examples than the other three interfaces (especially with few learning examples). The other three interfaces gave approximately equal results.

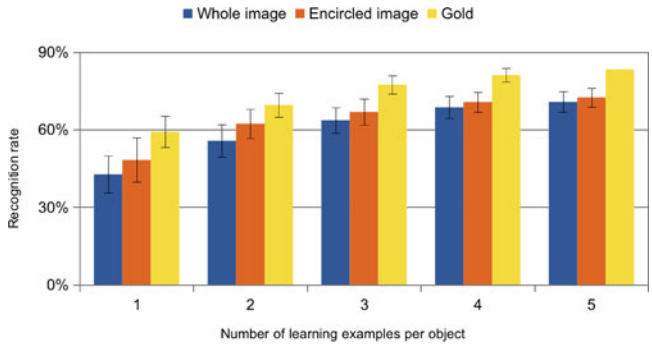

Fig. 21. Impact of encircling with the iPhone interface on the recognition rate. As we can see, this intuitive gesture allows us to improve the recognition rate, especially when the system is trained with very few learning examples.

we obtained significantly higher results with the iPhone interface than with the three other interfaces. Therefore, while the interface does not seem to have a very strong impact on the recognition of the big objects, interfaces, such as the iPhone interface, allow users to obtain a significantly higher recognition rate for small objects, especially with very few learning examples. Those results are coherent with the qualitative results presented earlier.

In the aforementioned tests, the whole image was used as an input. Thus, the encircling feature of the iPhone interface was not leveraged. We also investigated how encircling impacts the performance of the overall system. As we can see in Fig. 21, encircling with the iPhone allows us to improve the performance of the system, especially, when the system is trained with very few learning examples. In particular, we can see that the recognition rate is between 5% and 10% higher when trained with less than three encircled learning examples. Yet, we did not find any statistical difference here. Although the experimental environment was cluttered to reproduce a plausible environment, the background was still relatively plain in comparison with many real-world environments where this result would probably be even more important.

We also studied the impact of encircling with the Laser interface. However, we did not find any difference between the two conditions: whole or encircled images. We thus looked in detail at the images collected with this interface and found that

Fig. 22. While the feedback provided by the Laser interface allows users to make sure that the object is visible, it does not help them to realize how the object is actually perceived by the robot. (Left) Video game is almost entirely occluded by the table. (Center) Cluttered foreground in front of the poster of Zidane. (Right) Image of the magazine has been taken with an almost horizontal point of view.

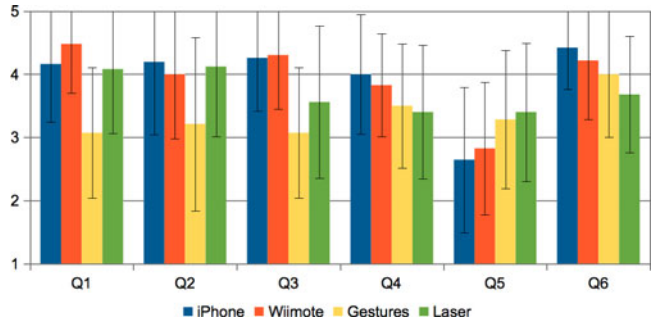

Fig. 23. *Usability:* Participants found the gestures interface significantly less intuitive and harder to use than the other interfaces. They also stated that the iPhone interface was overall more pleasant than the Laser interface. Q1: It was easy to learn how to use this interface. Q2: It was easy to move the robot. Q3: It was easy to make the robot look at an object. Q4: It was easy to interact with a robot. Q5: The robot was slow to react. Q6: Overall, it was pleasant to use this interface.

in many cases, the encircling was correct and should theoretically improve the results. Yet, in many other cases, the laser stroke cut the object, and therefore, the encircling actually led to a decrease in the quality of the learning examples (as shown in Fig. 10).

These results allow us to show that the interface plays an important role, allowing nonexpert users to robustly teach visual objects to a robot. In particular, we showed that the interface has a strong impact on the quality of the learning examples gathered by users. Indeed, we first showed that with simple interfaces, such as the Wiimote or the Gesture interfaces that do not provide any feedback to the users, only 50% of the learning examples collected by users can be considered "good." We also found that encircling improves the recognition rate. Furthermore, we showed that three examples provided by the iPhone interface allow us to obtain a higher recognition rate than eight examples collected with any other interface. These results are particularly important to us as we state that real-world users would probably want to give very few learning examples as it could quickly become a tedious task.

We also showed that specifically designed interfaces, such as the Laser and the iPhone interfaces, which provide the users a feedback of what the robot perceives, allow nonexpert users to ensure that the object they want to teach is actually visible. While it was expected that providing feedback to the users will help them to collect "apparently" better learning examples, it is very interesting to see that only the examples collected by the iPhone interface led to a very significant improvement of the overall performance of the learning system. Indeed, the Laser and the Wiimote interfaces gave a rather equal recognition rate. Thus, we can see that the kind of feedback of what the robot perceives also strongly influences the quality of the examples. More precisely, we think that while the Laser interface allows users to know whether an object is visible or not, it provides no information on how the object is actually perceived by the robot. For instance, as shown by the examples in Fig. 22, many examples were captured either far from the object (therefore, the object was very small in the picture), or the taught object was in the background, while other uninteresting objects were in the foreground, etc.

Thus, if one is interested in allowing nonexpert users to teach visual objects to a social robot by providing very few learning examples, we think that the interface should really be taken into consideration and specifically designed. In particular, as naive participants seem to have strong incorrect assumptions about a humanoid's visual apparatus, we argue that the design of the interface should help users to better understand what the robot perceives but should also drive them to pay attention to the learning examples they are collecting. For instance, the iPhone interface presents, on the screen of the device, the learning example that users encircle and provide to the robot. Thus, the interface naturally forces them to monitor the quality of the examples they collected.

C. Subjective Evaluation of the Usability and Game Experience

Fig. 23 presents the answers to the usability questionnaires. We performed a one-way ANOVA where the independent variable was the interface used and the dependent variable was the answer given in the questionnaires. We found statistical differences for the questions Q1 ($F_{3,103} = 6.35, p < 0.001$), Q2 ($F_{3,103} = 2.44, p < 0.05$), Q3 ($F_{3,103} = 6.41, p < 0.001$), and Q6 ($F_{3,103} = 3.38, p < 0.05$). The Tukey posthoc tests showed that the iPhone, Wiimote, and Laser interfaces were judged as easier to learn and more practical to move the robot than the Gesture interface. The users also stated that it was easier to make the robot look at an object with the iPhone and Wiimote interfaces. Furthermore, they also judged that overall the iPhone was significantly more pleasant to use than the Laser interface. In particular, during the experiments, we observed that the Gesture interfaces led to some misunderstanding while interacting, and therefore, participants tended to rush through the experiment.

Fig. 24 shows the results for the game part of the questionnaires. The only statistical difference was found for question Q1. We can see that the participants found the game easier when using interfaces based on mediator objects rather than with the gestures interfaces ($F_{3,103} = 5.17, p < 0.005$). The game was judged as entertaining by participants for all conditions. It is also interesting to note that the gesture's condition seems to improve the feeling of cooperation with the robot. Similarly, participants seemed to be more willing to play other robotic games with the gestures interface in the future than with the other conditions. However, no statistical difference was found for these results.

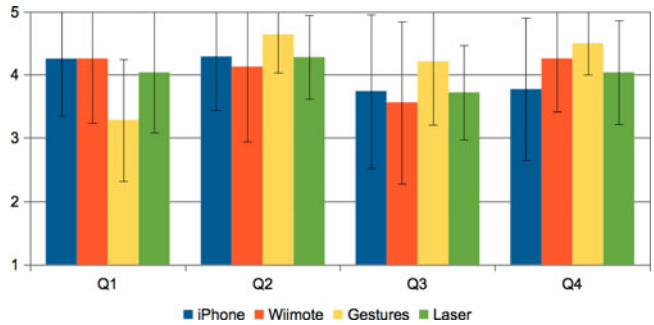

Fig. 24. *Robotic game:* Our robotic game was stated as entertaining by all participants. They found the game significantly harder with the gestures interfaces, but it increased the feeling of cooperation with the robot. Q1: Completing the game was easy. Q2: The game was entertaining. Q3: I felt like cooperating with the robot. Q4: I picture myself playing other robotic games in the future.

It is interesting to note that while the gestures interface was stated as being less usable than the other three interfaces, participants judged that the game was as entertaining with this interface as with the others. To us, this result can be explained by several factors. First, it is important to notice that the participants did not know whether they collected good learning examples or not; it did not influence the user's experience. For instance, users who collected only very bad learning examples could still think that they had successfully finished the game. Second, while interfaces, such as the iPhone interface, were specifically designed to help users collect good learning examples, it was probably more complicated than necessary for the robotic game. Indeed, users had to monitor the robot, the game interface, and the iPhone. Furthermore, it seemed that the interface should be as transparent as possible in order to allow users to entirely focus on the game. Finally, the gestures interface seemed to improve the user's feeling that the robot was cooperating with them. We think that this result could be explained by the fact that participants were closer to the robot and that they were trying different gestures to see how the robot reacted and therefore thereby determining which gestures were better understood. The bias introduced by the Wizard-of-Oz setup also led to a situation where the Wizard was adapting its behavior to the participants and, thus, was effectively cooperating with them. Although further studies should be carried out in this direction, our preliminary results about this seem to show that the gestures interface could be interesting if one tries to develop a simple robotic game.

As stated previously, we also timed the experience. However, we did not find any significant difference between the different interfaces. Furthermore, for all the aforementioned results, no significant differences was found between the participants from the science museum and the participants from the university campus. We also studied other sociological differences (age, gender, etc.) without finding any remarkable result.

VI. Discussion and Future Work

We have proposed an integrated system, which is based on a combination of advanced interfaces, visual perception, and machine learning methods, that allows nonexpert users to teach a robotic system how to recognize new visual objects. For experimental convenience, the robotic system was composed of a personal mobile robot and a remote computer which achieves offline signal processing and statistical learning. All the signal processing and statistical learning algorithms used are fast and incremental, and it is possible to use them online and onboard a mobile robot [54].

With this system, we have studied the impact of various interfaces based on mediator objects upon the efficiency of robot learning. We have shown that a well-designed interface, such as the iPhone, permits the collection of high-quality learning examples by nonexpert users in realistic conditions and outside of the lab. In particular, providing feedback about what the robot perceives allows nonexpert users to collect as good learning examples as expert users who are familiar with the robotic and visual recognition issues.

We have also shown that our interfaces, which are based on mediator objects, were judged intuitive and easy to use by participants. Users stated that they were more intuitive and more efficient to teach new visual objects to a robot than a direct transposition of a human-like interaction based on gestures. This result can be explained by the misconceptions held by nonexpert users about what the robot perceives visually. Artifact-based interfaces allow users to understand better what the robot perceives, and, thus, guide users into providing good learning examples.

In this paper, we also investigated how a robotic game can be used as a framework to conduct real world and large-scale user studies. We argue that such an approach allows the design of an experimental setup that engages and motivates users and justifies an *a priori* artificial task but also has a specific and reproducible protocol. We have also explored the concept and realization of the robotic game that raises interesting questions, especially since robotic games are a relatively unexplored research area [55], [56]. Our preliminary result seems to show that users were entertained by our games and were willing to play more robotic games. However, it seems that the game should be rather simple and the interface transparent to allow users to focus on the game play. Further studies should be conducted in these directions.

The experimental setup presented here was designed following some particular design choices. As stated previously, we chose to restrict ourselves to the use of the Nao robot without enhancing its onboard sensors. In our opinion, this robot was representative of some particular characteristics that made it a plausible candidate for the future of personal robotics such as having a humanoid shape, being relatively low cost and easy to integrate into a domestic environment both from a technological and cultural point of view.

This robot also had some limited sensors which constrained the interaction; however, this limitation is currently difficult to avoid in robots targeted for everyday use in the home. As long as robots and humans have different sensorimotor and cognitive apparatuses, robustness issues during interaction will be inevitable, especially if one tries to directly transpose the human-like interaction into the HRI domain.

Other approaches, such as smart environments or ubiquitous robotics, have also widely been explored in the literature [57], [58]. Using an omnidirectional camera or a set of fixed camera

on the ceiling would have changed some of our results, and, in particular, it would have probably facilitated the drawing of attention when using the gesture-based interface and, thus, improving its usability.

Nevertheless, we argue that the interaction problems, such as interpreting pointing gestures robustly, knowing what the robot can perceive, and making sure that the robot is looking at what the object users are showing, would still remain despite these possible enhancements. It is interesting to note that our interfaces could also be combined with external cameras if they are available. For instance, a touch gesture-based interface or a laser pointer interface has been combined with ceiling cameras to facilitate drawing a service robot's attention [15], [39]. Further experiments should be conducted both in this direction and with other types of robots to evaluate the impact of the robotic setup on our interfaces and on their perception by users.

In the system presented in this paper, the visual objects were automatically labeled, which was made possible by the game context. More precisely, as we used a predefined set of objects during our experiments, they could automatically be associated with a particular symbol provided by the game interface. Such symbolic labels can be easily and surely compared, which then allows a direct classification of the different visual examples. This is an important feature as the clustering of the learning examples permits the construction of a better statistical model of the visual object and, therefore, a better recognition. However, for more diverse kinds of interaction, we should provide the ability for the users to directly enter a word that they want to associate with the object. Pilot studies have shown that the user would prefer to use vocal words [59]. In future work, we will thus investigate the use of acoustic words associated with visual objects without using an "off-the-shelf" speech recognition system, as we argue that they still suffer from robustness problems when used on single words and in uncontrolled conditions. We will, in particular, investigate the role of the interface to improve the speech recognition, for instance, by displaying the N closest words to the users and allowing them to choose among those possibilities. We will also study how the interface could provide the ability for the humans to incrementally build the complete clusterization of the different learning examples through intuitive and transparent interactions and, thus, circumvent the issues raised by not using symbolic labels.

Finally, in the experiments described previously, we chose to only perform offline visual classification. Indeed, the experiments were already rather complex and time consuming for the users, and therefore, we decided not to include the search part in our robotic game to keep the experimental time acceptable for users facilitating the testing of many users. It will be interesting to evaluate our integrated system in a complete scenario including the search of the objects in order to study the entire interaction, and let users have feedback on the learning efficiency of the robot. The robot should itself assess the quality of the learning examples. In such a closed-loop scenario, it will be interesting to investigate how the robot could provide feedback to the user regarding the quality of the learning examples collected or how the robot could use active learning techniques to ask informative questions of the user [60].

APPENDIX
VIDEO LINKS

1) For a description of the interfaces based on mediator objects, see http://www.youtube.com/watch?v=vrMsaIj2SDM&list=UU1fhtoUjlb-FLnyMSrQOnMQ&index=4
2) For a description of the gesture-based interface, see http://www.youtube.com/watch?v=15GOCqXdgQg&list=UUlfhtoUjlb-FLnyMSrQOnMQ&index=3

ACKNOWLEDGMENT

The authors would like to thank H. Fogg for her comments and spelling corrections. They would also like to thank the reviewers for their insightful comments. They also thank J. Béchu for his implication in the development and the realisation of the different parts of the system.

REFERENCES

[1] B. Gates. (2007, Jan.). "A robot in every home," *Sci. Amer.* [Online]. Available: http://www.sciam.com/article.cfm?id=a-robot-in-every-home
[2] T. W. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robot. Auton. Syst.*, vol. 42, pp. 143–166, 2003.
[3] C. L. Nehaniv and K. Dautenhahn, Eds., *Imitation and Social Learning in Robots, Humans, and Animals: Behavioural, Social and Communicative Dimensions*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[4] A. Billard, S. Calinon, R. Dillmann, and S. Schaal, "Survey: Robot programming by demonstration," *Handbook of Robotics*. New York: Springer, 2008, ch. 59.
[5] A. L. Thomaz and C. Breazeal, "Teachable robots: Understanding human teaching behavior to build more effective robot learners," *Artif. Intell. J.*, vol. 172, pp. 716–737, 2008.
[6] A. Thomaz and C. Breazeal, "Robot learning via socially guided exploration," in *Proc. IEEE 6th Int. Conf. Dev. Learn.*, Jul. 2007, pp. 82–87.
[7] S. Calinon, F. Guenter, and A. Billard, "On learning, representing, and generalizing a task in a humanoid robot," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 37, no. 2, pp. 286–298, Apr. 2007.
[8] P. Miller, *Theories of Developmental Psychology*, 4th ed. New York: Worth, 2001.
[9] S. Calinon and A. G. Billard. (2007). What is the teacher's role in robot programming by demonstration? toward benchmarks for improved learning, *Interact. Stud. Spec. Issue Psychol. Benchmarks Human-Robot Interact.* [Online]. vol. 8, no. 3, pp. 441–464. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.73.2276
[10] M. Tomasello, M. Carpenter, J. Call, T. Behne, and H. Moll, "Understanding and sharing intentions: The origins of cultural cognition," *Behav. Brain Sci.*, vol. 28, no. 5, pp. 675–690, 2005.
[11] F. Kaplan and V. Hafner. (2004). The challenges of joint attention. *Proc. 4th. Int. Workshop Epigenetic Robotics* [Online]. pp. 67–74. Available: http://cogprints.org/4067/
[12] L. Steels and F. Kaplan, "Aibo's first words: The social learning of language and meaning," *Evol. Commun.*, vol. 4, no. 1, pp. 3–32, 2000.
[13] K. Dautenhahn and J. Saunders. (2011). *New Frontiers in Human-Robot Interaction* (Advances in Interaction Studies Series). Amsterdam, The Netherlands: John Benjamins. [Online]. Available: http://books.google.fr/books?id=_FlP3ZBhq6oC
[14] P.-Y. Oudeyer. (2011). "Developmental robotics," in *Encyclopedia of the Sciences of Learning* (Springer Reference Series), N. Seel, Ed. New York: Springer-Verlag [Online]. Available: http://hal.inria.fr/hal-00652123
[15] K. Ishii, S. Zhao, M. Inami, T. Igarashi, and M. Imai, "Designing laser gesture interface for robot control," in *Proc. 12th IFIP Conf. Human–Comput. Interact.*, 2009, pp. 479–492.
[16] 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," in *Proc. IEEE Int. Symp. Robots Human Interact. Commun.*, 2009, pp. 1066–1072.
[17] 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 *Proc. IEEE-RAS Int. Conf. Humanoid Robots*, 2009, pp. 391–398.

[18] P. Rouanet, P. Oudeyer, and D. Filliat, "Using mediator objects to easily and robustly teach visual objects to a robot," in *Proc. ACM SIGGRAPH Posters*, 2010.
[19] P. Rouanet, F. Danieau, and P.-Y. Oudeyer, "A robotic game to evaluate interfaces used to show and teach visual objects to a robot in real world condition," in *Proc. 6th Int. Conf. Human-Robot Interact.*, 2011, pp. 313–320.
[20] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, pp. 1470–1477.
[21] D. Filliat, "Interactive learning of visual topological navigation," in *Proc. IEEE Int. Conf. Intell. Robots Syst.*, 2008, pp. 248–254.
[22] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005.
[23] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.
[24] D. K. Roy, "Learning words from sights and sounds: A computational model" Ph.D. dissertation, Dept. Brain Cognitive Sci., Mass. Inst. Technol., Cambridge, 1999.
[25] C. Yu and D. H. Ballard, "A multimodal learning interface for grounding spoken language in sensory perceptions," *ACM Trans. Appl. Percept.*, vol. 1, pp. 57–80, 2004.
[26] L. Steels and T. Belpaeme, "Coordinating perceptually grounded categories through language: A case study for colour," *Behav. Brain Sci.*, vol. 28, no. 4, pp. 469–488, 2005.
[27] F. Kaplan, "Talking AIBO: First experimentation of verbal interactions with an autonomous four-legged robot," in *Proc. Learn. Behave: Interact. Agents CELETWENTE Workshop Lang. Technol.*, 2000, pp. 57–63.
[28] B. Scassellati, "Mechanisms of shared attention for a humanoid robot," presented at the Amer. Assoc. Artif. Intell., Fall Symp. Embodied Intell., Cambridge, MA, 1996.
[29] D. Perzanowski, A. C. Schultz, W. Adams, E. Marsh, and M. Bugajska, "Building a multimodal human-robot interface," *IEEE Intell. Syst.*, vol. 16, no. 1, pp. 16–21, Jan.–Feb. 2001.
[30] A. Haasch, S. Hohenner, S. Huwel, M. Kleinehagenbrock, S. Lang, I. Toptsis, G. Fink, J. Fritsch, B. Wrede, and G. Sagerer. (2004). "Biron—The bielefeld robot companion," in *Proc. Int. Workshop Adv. Serv. Robot.*, Stuttgart, Germany [Online]. pp. 27–32. Available: citeseer.ist.psu.edu/article/haasch04biron.html
[31] D. Roy, "Grounded spoken language acquisition: Experiments in word learning," *IEEE Trans. Multimedia*, vol. 5, no. 2, pp. 197–209, Jun. 2003.
[32] F. Kaplan, *Les machines apprivoisées comprendre les robots de loisir*. Paris, France: Vuibert, 2005.
[33] F. Lömker and G. Sagerer, "A multimodal system for object learning," in *Proc. 24th DAGM Symp. Pattern Recognit.*, 2002, pp. 490–497.
[34] H. Wersing, S. Kirstein, M. Götting, H. Brandl, M. Dunn, I. Mikhailova, C. Goerick, J. J. Steil, H. Ritter, and E. Körner, "A biologically motivated system for unconstrained online learning of visual objects," in *Proc. 16th Int. Conf. Artif. Neural Netw.*, 2006, vol. 2, pp. 508–517.
[35] C. C. Kemp, C. D. Anderson, H. Nguyen, A. J. Trevor, and Z. Xu, "A point-and-click interface for the real world: Laser designation of objects for mobile manipulation," in *Proc. 3rd ACM/IEEE Int. Conf. Human Robot Interact.*, 2008, pp. 241–248.
[36] K. Tsui, H. Yanco, D. Kontak, and L. Beliveau, "Development and evaluation of a flexible interface for a wheelchair mounted robotic arm," in *Proc. 3rd ACM/IEEE Int. Conf. Human Robot Interact.*, 2008, pp. 105–112.
[37] T. W. Fong, C. Thorpe, and B. Glass, "Pdadriver: A handheld system for remote driving," presented at the IEEE Int. Conf. Adv. Robot., Coimbra, Portugal, Jul. 2003.
[38] H. Kaymaz, K. Julie, A. Adams, and K. Kawamura, "PDA-based human-robotic interface," presented at the IEEE Int. Conf. Syst., Man Cybern., The Hague, Netherlands, Oct. 10–13, 2004.
[39] D. Sakamoto, K. Honda, M. Inami, and T. Igarashi, "Sketch and run: A stroke-based interface for home robots," in *Proc. 27th Int. Conf. Human Factors Comput. Syst.*, 2009, pp. 197–200.
[40] S. Lazebnik, C. Schmid, and J. Ponce. (2009, Nov.). "Spatial pyramid matching," in *Object Categorization: Computer and Human Vision Perspectives*, A. Leonardis, B. Schiele, S. J. Dickinson, and M. J. Tarr, Eds. Cambridge, U.K.: Cambridge Univ. Press [Online]. pp. 401–415. Available: http://hal.inria.fr/inria-00548647/en
[41] J. Wang, R. Cipolla, and H. Zha, "Vision-based global localization using a visual vocabulary," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2005, pp. 4230–4235.
[42] A. Angeli, D. Filliat, S. Doncieux, and J.-A. Meyer, "Real-time visual loop-closure detection," in *Proc. Int. Conf. Robot. Autom.*, 2008, pp. 1842–1847.
[43] 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.
[44] D. Nister and H. Stewenius. (2006). "Scalable recognition with a vocabulary tree," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Washington, DC. [Online]. pp. 2161–2168. Available: http://dx.doi.org/10.1109/CVPR.2006.264
[45] J. Ponce, T. L. Berg, M. Everingham, D. Forsyth, M. Hebert, S. Lazebnik, M. Marszałek, C. Schmid, C. Russell, A. Torralba, C. Williams, J. Zhang, and A. Zisserman. (2006). "Dataset issues in object recognition," in *Towards Category-Level Object Recognition*. New York: Springer-Verlag [Online]. pp. 29–48. Available: http://lear.inrialpes.fr/pubs/2006/PBEFHLMSRTWZZ06
[46] D. Schmalstieg, L. M. Encarnaçao, and Z. Szalavári, "Using transparent props for interaction with the virtual table," in *Proc. Symp. Interact. 3D Graph.*, 1999, pp. 147–153.
[47] M. Hachet, F. Dècle, S. Knödel, and P. Guitton. (2008). "Navidget for easy 3D camera positioning from 2D inputs," in *Proc. IEEE Symp. 3D User Interfaces* [Online]. pp. 83–89. Available: http://iparla.labri.fr/publications/2008/HDKG08
[48] W. Gander, G. H. Golub, and R. Strebel. (1994, Dec.). "Least-squares fitting of circles and ellipses," *BIT Numer. Math.* [Online]. vol. 34, no. 4, pp. 558–578. Available: http://dx.doi.org/10.1007/BF01934268
[49] M. L. Walters, S. N. Woods, K. L. Koay, and K. Dautenhahn, "Practical and methodological challenges in designing and conducting human-robot interaction studies," in *Proc. AISB Symp. Robot Compan. Hard Probl. Open Challenges Human-Robot Interact.*, Apr. 2005, pp. 110–119.
[50] H. A. Yanco, J. L. Drury, and J. Scholtz, "Beyond usability evaluation: analysis of human-robot interaction at a major robotics competition," *J. Hum.-Comput. Interact.*, vol. 19, no. 1, pp. 117–149, 2004.
[51] P. Rouanet, P.-Y. Oudeyer, and D. Filliat, "A study of three interfaces allowing non-expert users to teach new visual objects to a robot and their impact on learning efficiency," in *Proc. 5th ACM/IEEE Conf. Human-Robot Interact.*, 2010, pp. 185–186.
[52] D. R. Michael and S. L. Chen, *Serious Games: Games That Educate, Train, and Inform*. Cincinnati, OH: Muska & Lipman, 2005.
[53] A. Weiss, R. Bernhaupt, M. Lankes, and M. Tscheligi, "The USUS evaluation framework for human-robot interaction," in *Proc. Symp. New Front. Human-Robot Interact.*, 2009, pp. 89–110.
[54] D. Filliat, "A visual bag of words method for interactive qualitative localization and mapping," in *Proc. Int. Conf. Robot. Autom.*, 2007, pp. 3921–3926.
[55] A. G. Brooks, J. Gray, G. Hoffman, A. Lockerd, H. Lee, and C. Breazeal, "Robot's play: Interactive games with sociable machines," *Comput. Entertain.*, vol. 2, no. 3, pp. 10–10, 2004.
[56] M. Xin and E. Sharlin, "Playing games with robots—A method for evaluating human-robot interaction," in *Human Robot Interaction*. Vienna, Austria: Itech, 2007.
[57] T. Kim, S. Choi, and J. Kim, "Incorporation of a software robot and a mobile robot using a middle layer," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 37, no. 6, pp. 1342–1348, Nov. 2007.
[58] T. Seifried, M. Haller, S. Scott, F. Perteneder, C. Rendl, D. Sakamoto, and M. Inami, "CRISTAL: A collaborative home media and device controller based on a multi-touch display," in *Proc. ACM Int. Conf. Interact. Tabletops Surf.*, 2009, pp. 33–40.
[59] P. Rouanet and P.-Y. Oudeyer, "Exploring the use of a handheld device in language teaching human-robot interaction," in *Proc. AISB Workshop: New Front. Human-Robot Interact.*, 2009.
[60] M. Lopes and P.-Y. Oudeyer, "Active learning and intrinsically motivated exploration in robots: Advances and challenges," *IEEE Trans. Auton. Mental Dev.*, vol. 2, no. 2, pp. 65–69, Jun. 2010.

Pierre Rouanet received the Graduate's degree in computer science from the University Bordeaux I, Talence, France, and the Ph.D. degree with the FLOWing Epigenetic Robots and Systems Team, INRIA and Ensta-ParisTech, Talence.

His research topics include the study of the role of human–robot interfaces for teaching new words to a robot, the elaboration of new interfaces for efficient and intuitive joint attention and joint intention between the robot and the human (with touchpad, wiimote, and laser pointer), and large-scale user studies with nonspecialist users in order to show that interfaces can considerably increase learning efficiency.

Pierre-Yves Oudeyer (M'12) studied theoretical computer science with the Ecole Normale Supérieure de Lyon, Lyon, France, and received the Ph.D. degree in artificial intelligence from the University Paris VI, Paris, France.

He is currently responsible of the FLOWing Epigenetic Robots and Systems Team, INRIA and Ensta-ParisTech, Talence, France. He was a Permanent Researcher with Sony Computer Science Laboratory from 1999 to 2007. After working on computational models of language evolution, he is currently involved in research on developmental and social robotics, focusing on sensorimotor development, language acquisition, and life-long learning in robots. Strongly inspired by infant development, the mechanisms he studies include artificial curiosity, intrinsic motivation, the role of morphology in learning motor control, human–robot interfaces, joint attention and joint intentional understanding, and imitation learning. He has published a book, more than 80 papers in international journals and conferences, holds eight patents, and given several invited keynote lectures at international conferences.

Dr. Oudeyer has received several prizes for his research in developmental robotics and on the origins of language. In particular, he received the Laureate of the European Research Council Starting Grant EXPLORERS. He is the Editor of the IEEE Computational Intelligence Society Newsletter on autonomous mental development and Associate Editor of the IEEE TRANSACTIONS ON AUTONOMOUS MENTAL DEVELOPMENT, FRONTIERS IN NEUROROBOTICS and of the *International Journal of Social Robotics*.

Fabien Danieau received the M.S. degree in cognitics from the Ecole Nationale Supérieure de Cognitique, Bordeaux, France, in 2010. He is currently working toward the Ph.D. degree with Technicolor, Issy-les-Moulineaux, France, and with the VR4I Team, Institut national de recherche en informatique et en automatique, Rennes, France.

His research interests include haptic interactions, multimedia, and user experience.

David Filliat (M'07) received the Graduate's degree from the Ecole Polytechnique, Palaiseau, France, in 1997 and the Ph.D. degree in robotics from the Université Pierre et Marie Curie, Paris, France, in 2001.

After three years of research on robotic programs for the French armament procurement agency, he is currently a Professor with the Ecole Nationale Supérieure de Techniques Avancées ParisTech (Ensta-ParisTech), Palaiseau, and a member of the FLOWERS Research Team, INRIA and Ensta-ParisTech, Talence, France. His main research interests include visual perception, navigation, and learning in the frame of the developmental approach to autonomous mobile robotics.