Object learning through active exploration

Serena Ivaldi, Sao Mai Nguyen, Natalia Lyubova, Alain Droniou, Vincent Padois, David Filliat, Pierre-Yves Oudeyer, Olivier Sigaud

***Abstract*—This paper addresses the problem of active object learning by a humanoid child-like robot, using a developmental approach. We propose a cognitive architecture where the visual representation of the objects is built incrementally through active exploration. We present the design guidelines of the cognitive architecture, its main functionalities, and we outline the cognitive process of the robot by showing how it learns to recognize objects in a human-robot interaction scenario inspired by social parenting. The robot actively explores the objects through manipulation, driven by a combination of social guidance and intrinsic motivation. Besides the robotics and engineering achievements, our experiments replicate some observations about the coupling of vision and manipulation in infants, particularly how they focus on the most informative objects. We discuss the further benefits of our architecture, particularly how it can be improved and used to ground concepts.**

***Index Terms*—developmental robotics, active exploration, human-robot interaction**

I. INTRODUCTION

THE connection between motor exploration and learning object properties is a central question investigated by researchers both in human development and in developmental robotics [1], [2]. The coupling between perception and manipulation is evident during infants' development of motor abilities. The quality of manipulation is related to the learning process [3]: the information they acquire about objects guides their manual activities, while these activities provide them with additional information about the object properties [4], [5]. Infants carefully select their exploratory actions [6], [7] and social cues shape the way they learn about objects since their first year [8].

Researchers leverage these insights to make robots learn objects and concepts through active exploration and social interaction. Several factors have to be considered: for example, the representation of objects and sensorimotor couplings in a robotic-centric perspective [9], [10], [11], [12], the learning and exploration strategy [13], [14], and the way social guidance from a human teacher or caregiver can be blended with the aforementioned [15], [16]. The combination of these factors reflects in the robot's cognitive architecture. Although literature focusing on one or more aspects is rich and diverse (see [17] for a survey), integrated solutions are rare; even rarer are those where the robot builds its knowledge incrementally within a developmental approach. For example, in [18] the architecture is focused on interaction and emotions, while in [19] on cooperation and shared plans execution. In [20], [21] the architectures are based on high-level ontologies. Overall, those architectures are limited in two respects: first, they make considerable assumptions on the prior knowledge of the robot; second, they often segregate the development of the perceptual levels from that of the cognitive levels.

Fig. 1. The humanoid iCub in the experimental contexts: autonomous and socially-guided exploration.

In contrast, we believe that development plays an essential role for the realization of the global cognitive process, and that it should guide the design of the cognitive architecture of robots at many levels, from elementary vision and motor control to decision-making processes. The robot should ground its knowledge on low level multi-modal sensory information (visual, auditory and proprioceptive), and build it incrementally through experience. This idea has been put forward by the MACSi project[1].

In this paper, we present the design guidelines of the MACSi cognitive architecture, its main functionalities and the synergy of perceptual, motor and learning abilities. More focused descriptions of some parts of the architecture have been previously published by the authors: [22], [23] introduced the perceptual-motor coupling and the human-robot interaction functions, [24] the engagement system, [12] the vision tracking system, [25] the intrinsic motivation system. We describe the cognitive process of the robot by showing how it learns to recognize objects in a human-robot interaction scenario inspired by social parenting.

We report experiments where the iCub platform interacts with a human caregiver to learn to recognize objects. As an infant would do, our child robot actively explores its environment (Fig. 1), combining social guidance from a human "teacher" and intrinsic motivation [26], [27]. This combined strategy allows the robot to learn the properties of objects by

S. Ivaldi, A. Droniou, V. Padois and O. Sigaud are with Institut des Systèmes Intelligents et de Robotique, CNRS UMR 7222 & Université Pierre et Marie Curie, Paris, France. e-mail: serena.ivaldi@isir.upmc.fr.

S.M. Nguyen and P.-Y. Oudeyer are with Flowers Team, INRIA, Bordeaux - Sud-Ouest, France.

N. Lyubova and D. Filliat are with Flowers Team, ENSTA ParisTech, Paris, France.

This work was supported by the French ANR program (ANR-10-BLAN-0216) through Project MACSi, and partly by the European Commission, within the CoDyCo project (FP7-ICT-2011-9, No. 600716).

Manuscript received —; revised —.

[1]http://macsi.isir.upmc.fr

actively choosing the type of manipulation and concentrating its efforts on the most difficult (or the most informative) objects.

The paper is organized as follows. Section II outlines the cognitive architecture, particularly the motor and perceptual systems. Section III-A shows that manipulation has a direct impact on the way objects are perceived by the robot, justifying why the robot needs to have an efficient exploration strategy. Section III-B describes how social guidance and intrinsic motivation are combined for the active exploration for an object recognition task. In Section IV, we discuss the implications of the experimental results. In Section V we provide further insights on the perspectives of our work.

II. COGNITIVE ARCHITECTURE

Sensorimotor activities facilitate the emergence of intelligence during the interaction of a cognitive agent with the environment [28]. In robotics, the implementation of the cognitive process requires the edification of several perceptual, learning and motor modules that are typically integrated and executed concurrently on the robotic platform. The orchestration of such modules is defined within the design of the robot's cognitive architecture. As anticipated, the design of our architecture takes inspiration from developmental psychology and particularly from studies on infants development, which offers interesting lessons for developing embodied intelligent agents. Not only should the robot be able to develop its prospection and action space incrementally and autonomously, but it should be capable of operating in a social environment, profiting of humans to improve its knowledge.

A. *"Six lessons from infant development"[29]*

In [29], Smith & Gasser defined six fundamental properties that embodied intelligent agents should have and develop: multimodality, incremental development, physical interaction with the environment, exploration, social guidance and symbolic language acquisition. Our cognitive architecture meets the first five requirements and paves the way for the sixth.

- *Multimodality*: We rely on multiple overlapping sensory sources, e.g. auditory (microphone arrays), visual (cameras in the robot's eyes), somatosensory (proprioceptive, kinesthetic - joints encoders, inertial and force/torque sensors); as it is frequently done in robotics, we also include extrinsic sensory sources, e.g. external RGB-D cameras. The richness of perceptual information is a hallmark of humans, and a distinctive feature of the humanoid platform we use, iCub [30].
- *Incremental development*: Infants may have pre-wired circuits [31] but they are very premature in terms of knowledge and sensorimotor capabilities at their beginning. These capabilities mature during their development [32] as the result of a continuous and incremental learning process. To replicate such skills in our robot, the design of our cognitive architecture entails several autonomous and incremental learning processes at different levels. For example, we demonstrated how the robot can learn autonomously its visuo-motor representations in simple visual servoing tasks [23], and to recognize objects from observation and interaction [12], [33].
- *Physical interaction with the environment*: Intelligence requires the interplay between the human baby with his surrounding, i.e. people and objects. Crucially, interaction is essentially physical: babies exploit the physical support of their environment, manipulate objects, use physical contact as a means for learning from humans. Contact and touch are also the primary form of communication that a baby has with his mother and the dominant modality of objects' exploration (e.g. through mouthing) during the first months of life [34]. To make the robot interact physically with the environment and with people, in an autonomous or very little supervised way, the compliance of the platform must be suitably controlled. Put differently, the robot should be "safe". This requirement is met by the motor controllers developed in our architecture that exploit the sensory feedback to control the robot's forces during both intentional and accidental interactions [35], [36], [37].
- *Exploration*: Children explore their environment sometimes acting in a seemingly random and playful way. This non goal-directed exploration gives them opportunities to discover new problems and solutions. Open and inventive exploration in robotics can also unveil new action possibilities [27], [38], [39], [40]. In our architecture, we provide several tools to drive exploration, to combine it with intrinsic motivation and social guidance [22], [25]. Not only our motor primitives are safe so the robot can explore on its own (or minimally supervised by the human), but they are sufficiently numerous and assorted so the robot can perform simple and complex objects manipulations.
- *Social guidance*: Human babies can learn autonomously, but they learn the most during social interactions. In our system, the robot is able to follow and engage with the active caregiver [24]; the human in the loop can tutor the robot and influence the way it interacts with its environment.
- *Symbol and language acquisition*: Language is a shared and symbolic communication system, grounded on sensorimotor and social processes. In our architecture, we provide the base for grounding intermediate-level or high(er)-level concepts [41], for example the vision system categorizes and recognizes objects that the human interacting with the robot can label with their name. But we do not integrate or exploit language acquisition mechanisms yet.

The cognitive architecture is shown in Fig. 2: it consists of an integrated system orchestrating cognitive, perceptive, learning and control modules. All modules are tightly intertwined, and their numerous and different couplings enable the emergence of visuo-motor representations and cognitive loops. From the perceptual point of view, different sensory sources are used: external sensors and internal sensors embodied on the robotic platform. In the first group, we have microphones sound arrays, used to detect the direction of sound, and

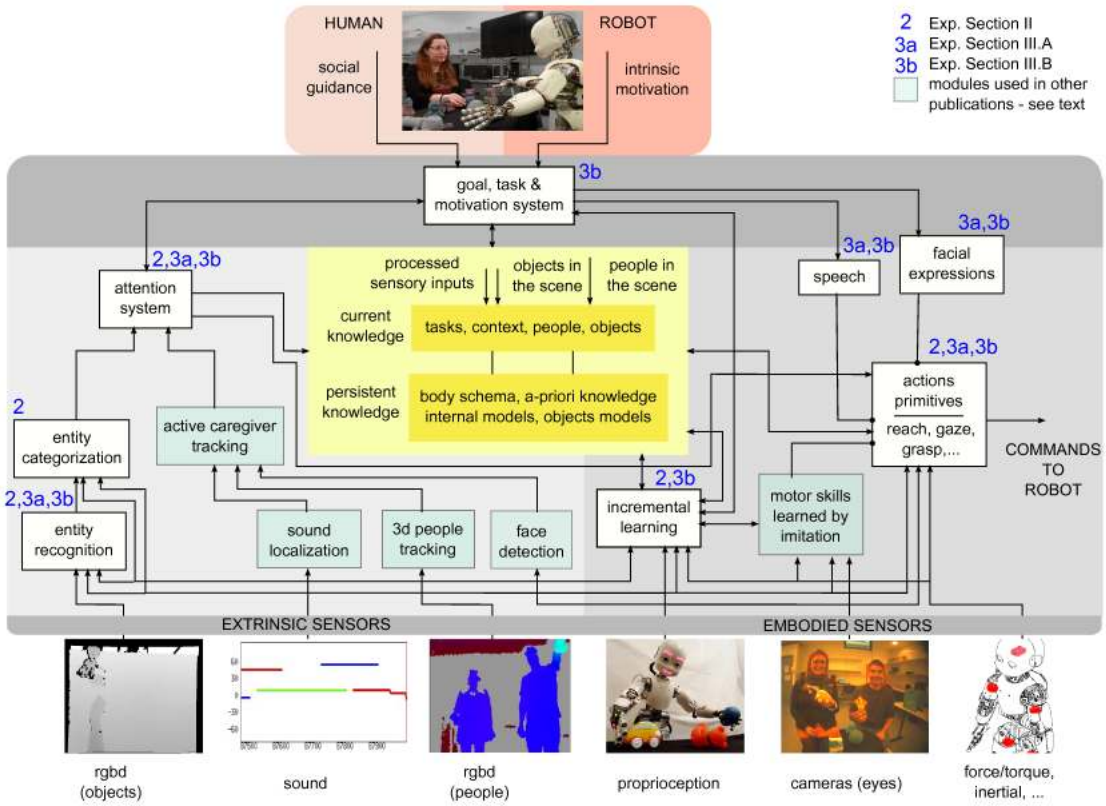


Fig. 2. The Cognitive Architecture of the MACSi Project: a functional description of its elementary modules. Human and robot are explicitly indicated as the two main "actors" influencing the behavior of the system: human can provide guidance (more generally, give commands), while the robot can act autonomously following its intrinsic motivation system. The pool of interconnected modules constitutes the learning and sensorimotor loops. A numeric legend is used to indicate the modules which are used in the experiments discussed in this paper (number n means "used in the experiments of Section n). Some modules, indicated by a colored background, are not used in the presented experiments, but have been used for the experiments of other publications. Precisely, the caregiver tracking modules are described in [22], [24], while the imitation learning module is described in [42].

RGB-D sensors, placed over a table to segment and detect objects or in front of the robot to detect interacting people (see Section II-D). In the second group, we have all the sensors embedded in the robotic platform, later described in Section II-E..

B. Decision and intrinsic motivation

The decision-making system is an autonomous process based on intrinsic motivation [13], [26], which combines social guidance with active exploration. The robot can exploit social guidance for bootstrapping or boosting its learning processes while exploring playfully or cooperating with humans to accomplish some tasks. This mechanism is crucial for the robot's cognitive system: given the huge space of visual and motor possibilities, selection and guidance are necessary to narrow down the exploration, and orient the robot towards "interesting" objects and events.

The expression *intrinsic motivation*, closely related to the concept of *curiosity*, was first used in psychology to describe the spontaneous attraction of humans toward different activities for the pleasure that they experience [43]. These mechanisms are crucial for humans to autonomously learn and discover new capabilities [44]. In robotics, they inspired the creation of meta-exploration mechanisms monitoring the evolution of learning performances [14], [27], [45], with heuristics defining the notion of interest used in an active learning framework [46], [47], [48].

The implementation of the intrinsic curiosity mechanism is done by the *Socially Guided Intrinsic Motivation with Active Choice of Teacher and Strategy* (SGIM-ACTS) algorithm [25] that combines interactive learning [49] and intrinsic motivation [50]. It achieves hierarchical active learning in a setting where multiple tasks and multiple learning strategies are available, thus instantiating Strategic Learning as formalized in [51]. It learns to complete different types of tasks by actively choosing which tasks/objects to focus on, and which learning strategy to adopt to learn local inverse and forward models between a task space and a state space. SGIM-ACTS is separated into two levels:

- A *Strategy and Task Space Exploration* level which decides actively which task/object to manipulate and which strategy to perform (*Select Task and Strategy*). To motivate its choice, it maps the task space in terms of interest level for each strategy (*Goal Interest Mapping*).
- A *State Space Exploration* level that explores according to the task-strategy couple chosen by the Strategy and Task Space Exploration level. With each chosen strategy, different samples state-task are generated to improve the estimation of the model. It finally returns the measure of error to the Strategy and Task Space Exploration level.

Details of SGIM-ACTS are reported in Appendix A, but more can be found in [25]. In Section III-B, we describe how SGIM-ACTS is used for an object recognition task in cooperation with a human teacher. Remarkably, the effective implementation of such mechanisms to address elementary challenges requires a tight coupling between the visual, cognitive, motor and learning modules, which is a novel feature of our architecture.

C. Action

Perceptive and cognitive modules are interfaced to the robot through an action/motor interface, which controls speech, facial expressions and upper-body movements. We define a set of actions that can be evoked by specifying their type π, e.g. *take, grasp*, and a variable list of parameters θ, e.g. the object name, its location, the type of grasp, *etc*. The k-th action π_k is generally defined as:

$$\pi_k(x, \theta_k) \ , \tag{1}$$

where $x \in \mathbb{R}^n$ is the initial state of the robot at the beginning of the movement, and $\theta \in \mathbb{R}^p$ is a vector defining the parameters characterizing the movement. The primitive definition entails both actions/skills which are learnt by demonstrations [42] or pre-defined parameterized motions. Interestingly, π can be an elementary action, such as an open-loop gaze reflex, a closed-loop reaching, but also a complex action (i.e. a combination/chain of multiple elementary actions), such as

$$\begin{aligned} \pi_k(x_0, \theta_k) = (\pi_i(x_0, \theta_{k,i}) \rightarrow \pi_j(x_1, \theta_{k,j}) \rightarrow \dots \\ \dots \rightarrow \pi_h(x_{N-1}, \theta_{k,h})) \ , \end{aligned} \tag{2}$$

where π_k is a chain of N actions, which are applied to the initial state x_0, and make the system state evolve into $x_1, \dots, x_N$. The elementary actions, with their basic parameters are:

- speak (θ: speech text)
- look (θ: (x, y, z), i.e. Cartesian coordinates of the point to fixate)
- grasp (θ: selected hand, grasp type, i.e. fingers joints intermediate and final configurations)
- reach (θ: selected arm, x, y, z, i.e. Cartesian coordinates of the point to reach with the end-effector, o, i.e. orientation of the end-effector when approaching the point)

More complex actions (without specifying their numerous parameters, but just describing the sequence[2]) are:

- take (reach and grasp)
- lift (upward movement)
- rotate (take, lift, reach the table with a rotated orientation, release - open the hand)
- push (reach the target from one side, push by moving the hand horizontally, then withdraw the hand)
- put-on (take, lift, reach the target from the top and release)
- throw (take, lift, release)
- observe (take, lift, move and rotate the hand several times, to observe an in-hand object)
- give (take, lift, reach the partner and release)

If unpredictable events[3] occur during the execution of an action, for example an unsuccessful grasp or a potentially harmful contact with the environment, one or more autonomous reflexes are triggered. These reflexes are pre-coded sequences of actions that may interrupt or change the execution of the current action or task. Overall, our action interface is quite rich in terms of repertoire of actions, because besides elementary actions (such as in [19]) we provide the robot with more complex actions for a wider exploration capability. It also has coupling with the learning modules, so as to provide reproduction of trajectories learnt by demonstration, such as in [52]. Differently from [53], we do not integrate language processing for letting the human define on-line new sequences of actions, because this matter is outside the scope of our project.

D. Visual perception

The perceptual system of the robot combines several sensory sources in order to detect the caregivers and perceive its environment. The primary source for object detection is a RGB-D sensor placed over the area where the interaction with objects and caregivers takes place.

The object learning and recognition module has been designed with the constraints of developmental robotics in mind. It uses minimal prior knowledge of the environment: in particular it is able to incrementally learn robot, caregiver hands and object appearance during interaction with caregivers and objects without complementary supervision. The system has been described in details in [12], [54]. A short overview is given here to complement the architecture presentation.

All information about the visual scene is incrementally acquired as illustrated in Fig. 3. The main processing steps include the detection of physical entities in the visual space as proto-objects, learning their appearance, and categorizing them into objects, robot parts or human parts.

At the first stage of our system the visual scene is segmented into *proto-objects* [55] that correspond to units of visual attention defined from coherent motion and appearance. Assuming that the visual attention of the robot is mostly attracted by motion, proto-object detection starts from optical flow estimation, while ignoring the regions of the scene that are far away according to the constraints of the robot's workspace. Then, the Shi and Tomasi tracker [56] is used to extract features inside moving regions and to group them based on their relative motion and distance. Each cluster of coherently moving points is associated with one proto-object and its contour is defined according to the variation of depth. Each proto-object is therefore tracked across frames and finally identified as an already known or a new entity.

Each proto-object appearance is incrementally analyzed by extracting low-level visual features and grouping them into a hierarchical representation. As a basis of the feature hierarchy

[2]More details can be found in the online documentation of the code: http://chronos.isir.upmc.fr/~ivaldi/macsi/doc/group__actionsServer.html.

[3]These events are usually captured by the sensors embedded in the robot. For example, we threshold the external forces at the end-effectors, estimated thanks to the proximal force/torque sensors [35], to detect potentially harmful contacts with the table.

we use SURF points [57] and color of superpixels [58] obtained by segmenting the scene into regions of similar adjacent pixels. These low-level features are grouped into pairs and triples incorporating local geometry and called mid-features. Both low- and mid-level features are quantized into dictionaries of visual words. The Bag of Visual Words approach with incremental dictionaries [59] is used to characterize the appearance of entities from different viewpoints that we call *views*. Views are encoded by the occurrence frequency of extracted mid-features. An overall entity appearance is characterized by a multi-view model constructed by tracking an entity across frames and collecting its views occurrence frequency.

Besides tracking, the association of the current view to an entity can also be based on appearance recognition when an object appears in the field of view. In this case, appearance-based *view* recognition is performed first, using all extracted mid-features to participate in a voting procedure that uses the TF-IDF (Term-Frequency - Inverse Document Frequency) [60] and a maximum likelihood approach. If the recognition likelihood is high, the view is identified as the most probable among already known views; otherwise, a new view is created. Then, appearance-based *entity* recognition is performed using the same approach based on the occurrence statistics of views among known entities.

During experiments on interactive object exploration, objects are often grasped and therefore move together with a human or a robot hand. Thus, our approach performs a double-check recognition [12] to identify simultaneously moving connected entities, so that each segmented proto-object is recognized either as a single view or several connected views, where each view corresponds to one entity.

Finally, all physical entities are classified into the following categories: *robot parts*, *human parts* or *manipulable objects* (see Fig. 4). The categorization method is based on the mutual information between the sensory data and proprioception [61] and on statistics on the motion of physical entities. Among the remaining entities, we assume that each object moves only when it is connected to another entity, either a robot or a human, and each object is static and independent of robot actions when it is single. Thus, the object category is identified from the statistics on its simultaneous motion with robot and human parts.

Using the ability to categorize entities, the models of objects previously constructed during their observation can be improved during robot interactive actions (Fig. 4). Since the manipulated object does not change during the robot action, its corresponding model can be updated with recognized views connected to the robot hand or with new views created from the features that do not belong to the robot hand. The updates with recognized views reduce noise in object models, while the updates with new views allow the robot to accumulate views corresponding to unseen perspectives of the objects. Experiments in section III A illustrate this capacity. Remarkably, this approach is robust with respect to partial occlusions of the entities (see Fig. 3) and particularly to the numerous visual appearances that the hand can assume when it interacts with the objects, because the continuous collection of views

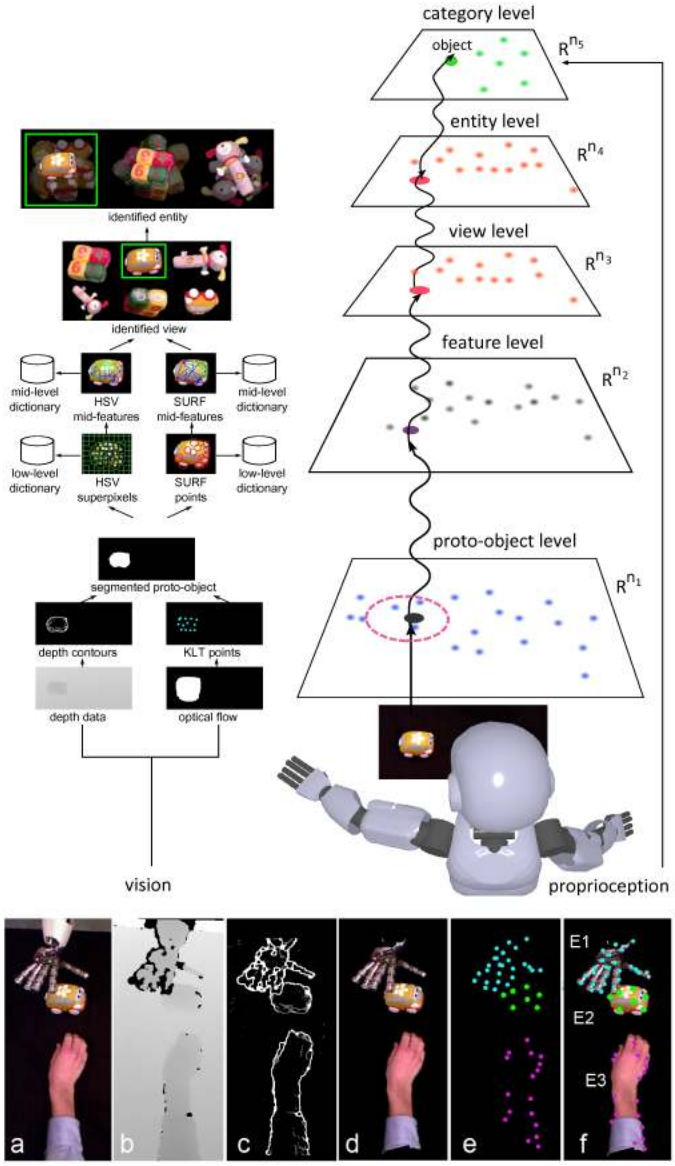


Fig. 3. The visual information is processed through a hierarchy of layers, which elaborate the camera images to extract the entities in the scene. The proprioceptive information from the robot is used for categorizing the entities - see Fig. 4 and text. The bottom part of the image shows the entity identification during human-object-robot interaction: (a) the camera image and (b) the depth map are retrieved from the RGB-D sensor; (c) contour extraction and (d) segmentation are fed to the processing system; (e) features are extracted and combined with the prior information (e.g. features and information extracted from the previous frames - see text) so as to determine (f) the entities in the scene.

contributes to the creation of a better model for recognizing it (see Fig. 4).[4]

Besides this unsupervised approach, a second RGB-D camera and the robot cameras are also used to detect the presence of people. In particular, the "active" caregiver, i.e. the human partner who gives feedbacks and instructions to the robot, is tracked through a multimodal approach, combining 3d

[4]To improve the performance of the visual system in correctly distinguishing the hand from the objects when they are first introduced in the visual field by the human, we start each experiment by showing the hand first to the robot - so the robot can gather enough data to build a good model of it.

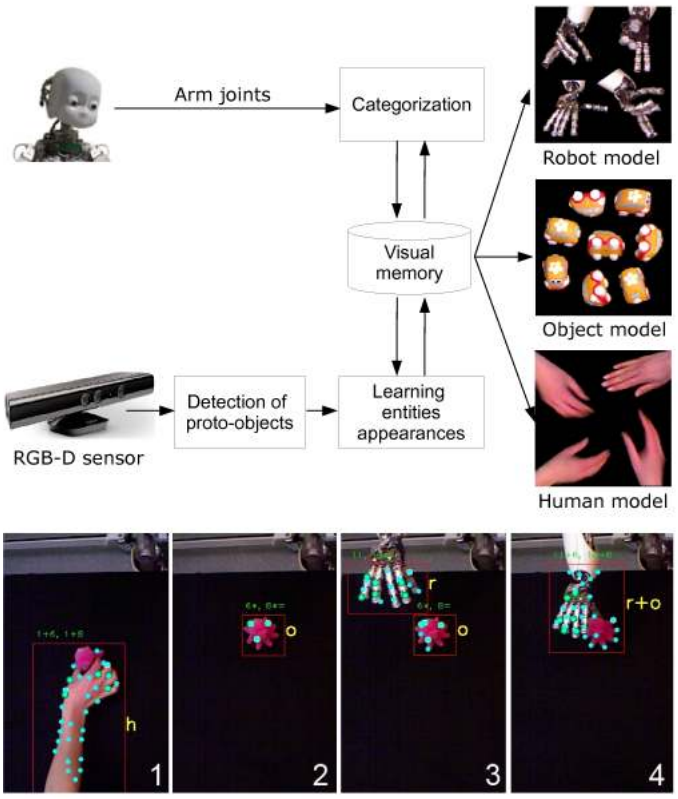


Fig. 4. Procedure for categorizing the entities in the images. The visual information is compared with the robot's proprioceptive information. The bottom part of the image shows the categorization of the entities identified in the images during a sequence of human-object-robot interaction. (1) The human shows an octopus toy to the robot: the *human hand*, moving during the action, is recognized (label *h*). (2) The octopus is the only inactive entity on the table, hence it is recognized as an *object* (label *o*). (3) The robot starts moving its hand to approach the octopus: two entities are recognized, and correctly identified as *robot hand* and *object* (labels *r* and *o*). (4) The robot grasps the octopus, so the two entities corresponding to the *robot hand* and the *object* are connected: the vision processing system is able to recognize this particular situation (labels *r+o*).

human estimation (via skeleton recognition) and direction of sound source. The robot gazes at the estimated pose of the active partner, arousing a natural eye-to-eye contact. More information about this engagement system can be found in [24], [22].

Perceptive modules communicate to the current/episodic knowledge of the robot. This module collects the processed perceptual information, such as the objects in the scene and their main features and properties retrieved by the vision modules (e.g. 3D position and orientation), the people interacting with the robot and their location, the current task etc. In a sense, it could be also seen as a sort of simplified "ego-sphere", more similar to the one of [62] (quoting the authors: "*a fast, dynamic, asynchronous storage of object positions and orientations*") than the Sensory-Ego-Sphere [63], [64].

E. Implementation and robotics setup

Our experiments were carried out with iCub, a 53 Degrees Of Freedom (DOF) humanoid robot shaped as a child [30], using the upper body of the robot (head, torso, arms and hands - totally 41 DOF). The proprioceptive sensors, the inertial and the proximal force/torque sensors embedded on the platform, combined with its kinematics and dynamics modeling, are fed to numerous modules of the architecture: modules for controlling gaze, posture and reaching movements [65], modules for controlling the whole-body dynamics [35], the robot compliance and its contact forces [36], modules for learning incrementally the visuo-motor models of the robot [23], the vision modules recognizing the robot's self-body in the visual space [33], the basic modules for speech and gaze tracking [24], just to cite a few.[5] The architecture is implemented as a set of concurrent modules exchanging information with the robot thanks to the YARP middleware [66]. Some modules, developed in ROS [67], communicate bidirectionally with YARP thanks to a simple bridge between the two middlewares.[6] Besides the basic modules necessary for starting the robot, the experiments described in the following section involved a number of modules developed within the MACSi Consortium, dedicated to the visual processing, the action interfaces and the curiosity system. Overall, each experiment required more than 20 modules, executed concurrently on 4 standard desktop computers (i5/i7 2.7GHz, 4/6GB RAM) connected through YARP. Details about this cluster can be found in the wiki pages of ISIR's iCub, while software details can be found in MACSi's online software documentation.[7] Each experiment took a variable time between 1-2 hours (experiments in Section III-B) to 6-7 hours (experiment in Section III-A). Fig. 5 shows the experimental setup: the human, the robot, the table and the RGB-D sensor. The calibration procedure required to match the RGB-D data within the robot reference frame is detailed in [22].[8]

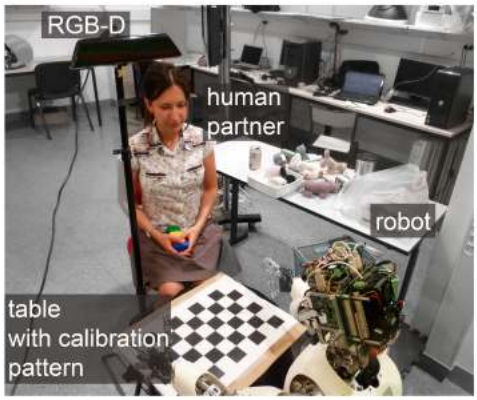


Fig. 5. The experimental setup.

III. EXPERIMENTS

In this section we discuss two sets of experiments performed with iCub (see Figure 1) that tackle some questions related to

[5]We did not integrate yet speech recognition; for speech production we use the standard *iSpeak* module from iCub's main software, based on a simple text-to-speech program called Festival.

[6]http://wiki.icub.org/wiki/UPMC_iCub_project/YARP_ROS_bridge

[7]Cluster: http://wiki.icub.org/wiki/UPMC_iCub_project/MACSi_cluster. Software: http://chronos.isir.upmc.fr/~ivaldi/macsi/doc/

[8]For more details about the calibration procedure and the software, see http://chronos.isir.upmc.fr/~ivaldi/macsi/doc/perception_table.html

the coupling of learning, vision and manipulation.

In Section III-A we show that some objects are more informative than others in terms of visual appearance, and that the robot needs to manipulate actively the object to gather all its views. Manipulation causes changes in the way the objects appear to the robot. Fig. 6 shows the outcome of two different sequences of actions applied to a yellow car and a blue train: some in-hand rotations and a side-push. The rotating action is more interesting and informative, since it causes the object to unveil hidden parts: more views can be then associated to this object, and this should help the robot to build a better model of the object. This is quite obvious for a human teacher providing the robot with new objects to learn, and commanding the robot to manipulate the objects: to boost the robot learning, a "good" teacher will always ask the robot to perform the maximally informative actions (i.e. the actions causing more changes in the objects' appearance, unveiling the objects' physical properties or their affordances). Similarly, the robot autonomous exploration should be guided by the will to gather more information about the objects, just like infants spend more time manipulating objects they are more willing to learn [3], [68]. In the robot, this curiosity process is implemented through the intrinsic motivation system [25].

In Section III-B we show how the intrinsic motivation system can be used to explore objects in a human-robot "parenting" scenario, to effectively choose how to manipulate the objects and which objects need to be explored the most. We also show how this exploratory mechanism can cope with "good" and "bad" teachers (i.e. naive partners) in the object recognition task. Incidentally, our approach has much in common with experimental observations about the way infants learn to recognize objects through feature tracking across actions and social cues [8].

A. Exploring objects through manipulation

When the robot exploration is driven by its intrinsic motivation, the curiosity mechanism should focus on the objects that are the most informative and difficult to recognize. To check this intuition, we performed a set of experiments where the robot interacted repeatedly with several different objects performing the very same actions. The goal was to highlight the objects that frequently changed their visual representation as a consequence of the action that was exerted on them. We took 14[9] different objects, shown in Fig. 8. The objects are either children toys, belonging to the set of "iCub's toys" (colorful objects that can be easily manipulated, grasped), or arbitrary items of the laboratory. Every object was pushed 30 times across the robot's workspace. Some examples of *pre/post* images (i.e. images of the object before and after the action occurs) are shown in Fig. 7. In some cases (7a), though pushing the object produced a change in its pose on the table, it did not substantially change the object's appearance to the visual system (for example it simply slid on the table surface, thus the main entity view attributed by the vision system did not change). In others (7b), pushing made the objects fall on their side, revealing a new or simply a different view that was added to the object model. Between each *pre/post* image, we measured the displacement of the object in a robot-centered reference frame and the change in the main (i.e. the most probable) view of the object. The results are shown in Fig. 9. As the median displacement in the $z-$axis is negligible, it is variable for the $x-$ and $y-$ axis. Intuitively, objects that have a cylindrical or spherical shape are able to "roll", thus move longer than others on the $y-$ axis (which is the axis perpendicular to the pushing performed by the robot). The emergence of the "rolling" property is discussed later in Section IV. Right now, we focus on the view-change. From the histograms (9b), we can see that the amount of views collected for each object is variable, and basically depends on the object physical and aesthetic properties. For example, the red bear had few different views, and there was basically no change in its appearance produced by the action ($<5\%$). Conversely, objects like the train, the chicken and the patchwork of plastic cubes changed very much their appearance, and on average a push provoked changes in their appearance ($>60\%$). Objects like the gray dog or the bottle had many views, but the push did not frequently alter their visual features (20-30%).

Overall, these results confirm two important observations. First, objects that are more complex and faceted require more exploration than those uniform in shape and color. Second, manipulation can change the visual appearance of the objects to the vision system, and this exciting outcome depends on the action and the object properties.

Consequently, an autonomous agent/robot that has no a priori knowledge of the surrounding objects, must necessarily interact with them to learn their properties. Intuitively the more the objects are complex, the more time it will spend on them, doing manipulation which can bring more information.

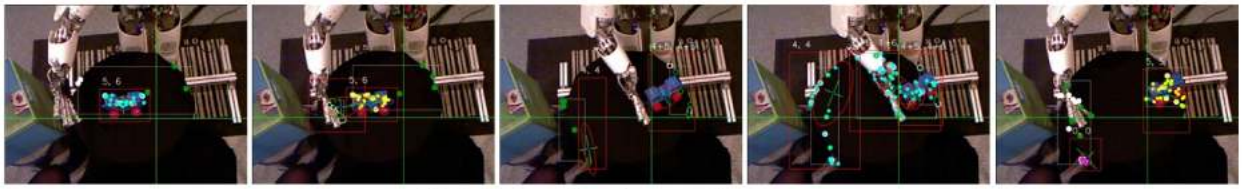
B. Exploration by social guidance and intrinsic motivation

As we discussed in Section II-A, the cognitive robot should engage in active exploration, driven by both social guidance [69], [70] and its own curiosity. Fig. 10 illustrates the human-robot interaction scenario and introduces the active learning strategy exploration. The robot's task is to learn to recognize different objects. At each decision step the exploration algorithm determines the triple *(object, action, actor)*, that is the object to explore, the action to perform on it, and the actor who is going to do the action. The robot can decide to do the action by itself, or it can ask the human to do it.[10] In the first case, the robot can perform one of the action primitives in its repertoire (Section II-C), for example a side push (which consists in approaching the object on its side, pushing and withdrawing the hand from the object) or a throwing (which

[9]It must be noted that the number of objects chosen for this study is not a limitation per se. The vision processing system learns from scratch and in an incremental fashion. It keeps recognizing previously shown entities, identifying new ones, adding more views to the entities, *etc.* (see [12] for more insights on the vocabulary growth and the vision performances). Since its knowledge is built incrementally, more objects could be easily added to the study.

[10]In both cases, the robot communicates its intent to the human through its speech synthesis system. For example it may say "*Now I push the bear*" or "*Show me the bear*". Some samples of communication can be found in [22], whereas videos giving a taste of the experiments can be seen in http://youtu.be/cN9eBaUpqWE and http://youtu.be/J7qfdWNe4uk.

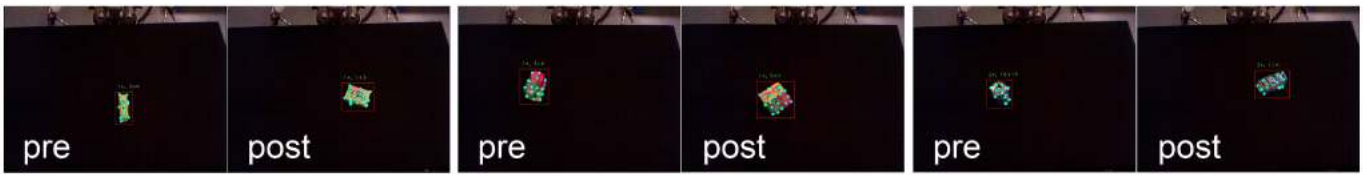
(a) Observing a grasped object.



(b) Pushing an object that can roll, e.g. a car.

Fig. 6. Combining action and perception. (a) The robot grasps an object and rotates it in its hand to take more views of it, and unveil hidden or yet unexplored parts. (b) The robot pushes an object to produce a change in its status. The vision system tracks the object: since the car rolls, the appearance of the object does not essentially change, only its location on the table.

(a) Example of objects views which are not changed by a *push-left* action. From left to right: red bear, white bottle, chicken.



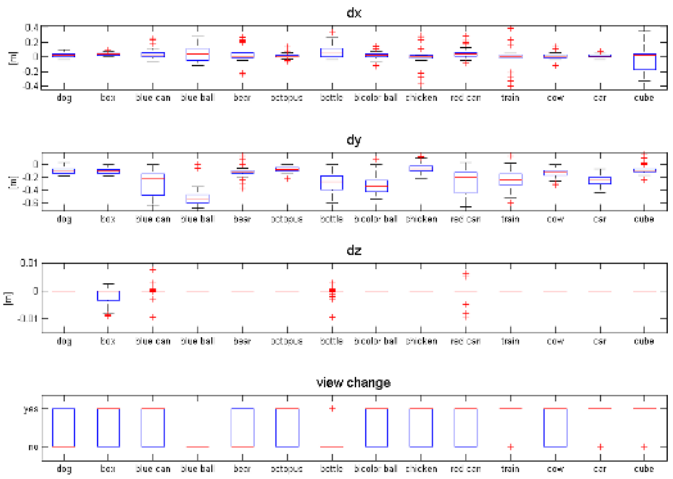
(b) Examples of object views which are changed by a *push-left* action. From left to right: green box, cubes, blue soda can.

Fig. 7. Some objects of Fig. 8 before (*pre*) and after (*post*) a *push-left* action. Though pushing an object produces a change in the object's location on the table, its global appearance may not necessarily change. In (a) the main view of the object does not change as an effect of the action for the red bear (*id*=2, *main view*=12), the white bottle (3,19) and the chicken toy (1,4). In (b) the main view of the objects changes, because pushing makes them fall on another side. In this case, the vision system adds a new view -or recognize a previously seen view- to the entity associated to the objects, respectively the green box (1,3 $\rightarrow$ 1,1), the patchwork of plastic colored cubes (1,3 $\rightarrow$ 1,5) - the entity id for the green box and the colored cubes is the same because the images were taken during two different experiments -, the blue soda can (2,18 $\rightarrow$ 2,11).

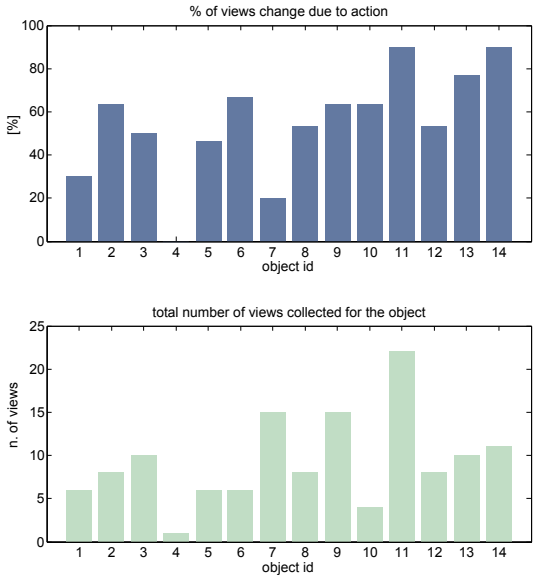
Fig. 8. The 14 objects used to study the effect of a *push-left* action. 1) gray dog 2) green box 3) blue soda can 4) blue ball 5) red bear 6) blue octopus 7) white bottle 8) blue/violet sphere 9) chicken 10) red soda can 11) yellow train 12) cow 13) yellow car 14) a patchwork of four plastic colored cubes.

consists in taking the object, lifting it and dropping it on the table - a "controlled" throwing which avoids the object to fall outside the visible space of the camera). In the second case, the robot can either ask the human to present a new object, or ask the human to manipulate the object, for example moving it on the table.

The idea is that exploratory and social motives can maintain the robot in a continuous state of "excitation" and acquisition of new stimuli, which nourish its learning process. The effect of the social drive however can be different. A "good" teacher can help consistently the robot during its learning process, for example showing different objects and presenting them in an informative way (e.g. showing all its sides). A "bad" teacher presenting objects the same way all times can be little helpful for the robot. However, this is particularly true for naive subjects, i.e. human partners that have no prior experience with the robot or do not know what the goal of the interaction process is. The intrinsic motivation system can counter-balance the social drive, by making the robot choose at each time the best exploration strategy for his task that can incorporate or not the human input. This intuition is evident in the following



(a) The first three plots (from the top) show the median value and the first and third quartiles of the *x*-,*y*- and *z*-axis displacement of the objects after a *push-left*. The fourth plot shows the changes of view.



(b) Effect of the *push-left* action on the perception of objects. Top: the view changes due to the action, i.e. the quota of actions that produced a change in the main view of the object. Bottom: the total number of views collected for each object.

Fig. 9. The effect of the *push-left* action on the 14 different objects (Fig. 8). We collected 30 samples (i.e. results of 30 different pushes) for each object. We measured the displacement of the object before and after the action (*pre* and *post*, cf. Fig. 7) in the *x*-, *y*- and *z*-axis in a robot referenced coordinates frame. The median displacement in the *z*-axis is evidently negligible. The histograms show the effect of the action in terms of changes in the perception of the object.

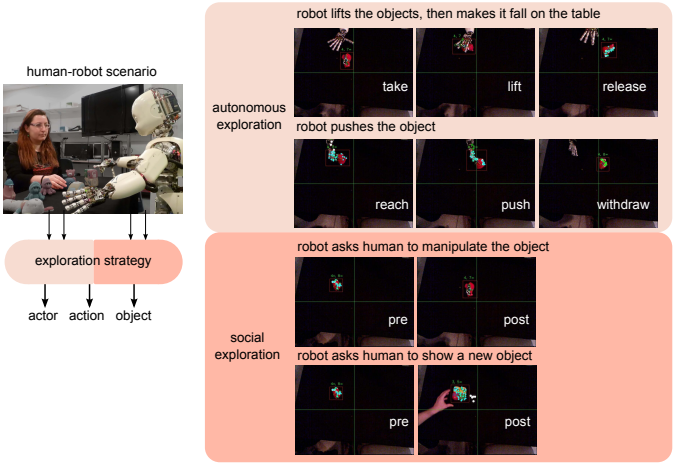


Fig. 10. The active exploration strategy in the human-robot interaction scenario: at each decision step, the exploration algorithm determines the triple *(object, action, actor)*, that is the object to explore, the action to perform on it, and the actor who is going to do the action (either robot or human). The choice of the triple can be done by the human caregiver if the robot is completely passive as in [71], or by the intrinsic motivation system if the robot is active. *Top:* when the exploration is active, the robot decides to manipulate the object, and can choose for example between two different actions: pushing or throwing the object. A pushing sequence consists in reaching the object on one side then pushing it in a direction parallel to the *y*-axis of the robot (e.g. towards the left with the right arm). A throwing sequence consists in reaching the object from the top, lifting it then opening the hand at once to release the object. The object drop is assimilated to a throwing, but of course it is more controlled and reasonably keeps the fallen object in the robot's workspace. As objects fall, their appearance changes quite unpredictably. *Bottom:* during active exploration, the robot can decide to ask the human to show a new object, or to manipulate the current one. In the first case, the human simply intervenes to change the object on the table. In the second case, the human can act on the object. If the human is a "good teacher", he can change radically the appearance of the object, for example by flipping it or putting it on one side: this action is beneficial for the robot, because the robot can have a new experience of the object, take more views, etc. If the human is a "bad teacher", he simply moves the object without caring to change its appearance to the robot.

experiment, where the intrinsic motivation system copes with the two types of teachers in an object recognition task.

We hereby present experimental results showing how the intrinsic motivation system incorporates social teaching to autonomously learn to recognize different objects. The human-robot scenario is presented in Fig. 10: the human cooperates with the robot by showing and manipulating some objects to learn, upon the robot request, while the robot manipulates the objects autonomously. We chose five objects among the ones of Fig. 8, namely the gray dog-like toy, the blue/violet ball, the red bear, the yellow car and the patchwork of yellow-red-green plastic cubes. With this choice, we mixed items that are easy and difficult to recognize because of their color and shape properties. Fig. 16b shows different views of the chosen objects. For example, the gray dog is easy to recognize because its color and shape are quite different from the others; the yellow car and the blue/violet ball are easier to distinguish in term of colors, however depending on their current orientation and pose on the table, their top view from the RGB-D camera may appear different in terms of color features or dimensions[11]; the patchwork of colored plastic cubes is the trickiest object to recognize, because its side views change the perceived dimension of the object (see Fig. 16b), and because of the different colors of the four cubes, the features of the global object can be confused with the ones of the car and the bear. In summary, we expect

[11]From a top view the ball may appear as a blue or as a violet circle, or as a mixed blue/violet circle. The yellow car appears as a big yellow rectangle, but with different sizes depending on the way the object lies on the table. In particular, it may be showing or not some characteristic features like the wheels or the toy decorations. A sample of the possible views of the car is shown in Fig. 16b.

toys like car and plastic cubes to arouse the interest of the cognitive agent: because of their rich "perceptive affordance", interaction can reveal their numerous views. An experiment consists of a sequence of interactions with the human and the objects: the number of interactions and their type change depending on the exploration strategy, the learning progress and the type of teacher.[12] At each decision step, the curiosity system determines the triple *(object, action, actor)*, that is the object to explore and one among the following action/actor possibilities:

- the robot pushes or lift/throws the object
- the human manipulates/shows the object

The teacher first labels all the objects. [13] Since the robot asks the teacher to switch the objects, it is aware at each step of the object it is currently manipulating, so it can collect views for the specified object label. Views are collected when the objects are immobile on the table, hence before (*pre*) and after (*post*) the actions: this is intentionally done to avoid collecting views of the moving entity, which would be demanded to future experiments. The vision processing system, of course, is never interrupted or inactive, because it needs to keep identifying entities and tracking their movement in the scene, estimate their location, etc. At the action completion (*post*), when the object is generally immobile on the table (notably, in a different pose), the vision learning system is triggered and views are collected again. The robot tests which object label/category it associates with the new object image, computes a confidence measure on its capability to recognize the object, and sends the evaluation results to the curiosity system. Depending on the progress, the curiosity system decides the next action to trigger. The learning progress is evaluated on the classification accuracy of the system on an image database (see Appendix B), made up of 64 images of each object in different positions and orientations, as shown in Fig. 16. Details of the learning algorithm, called SGIM-ACTS, are reported in Appendix A. Experiments were performed with two different types of teacher: a "good" teacher that we call "unbiased", which manipulates the objects at each time the robot asks, simply translating the object or showing a different side of it; and a "bad" teacher that we call "biased", which does not manipulate the objects when asked (i.e. it does not alter their appearance) and when asked to show a new object, always shows the same side. To have a fair comparison about the effectiveness of the curiosity system, we compared its learning progress with the one produced by a random exploration strategy, where the object and action to perform are picked up randomly.

We present and compare one exemplifying experiment for each of the four aforementioned conditions. Fig. 11 shows the number of images of the evaluation database which are correctly recognized over time. Fig. 12 detail the learning progress and the decision of the exploration strategies over time: each graph shows the progress in the f-measure (i.e. the harmonic mean of precision and recall [72]) for the five objects during time, while the bottom rows represent with a color code the chosen object and action at each decision time. The three actions are labeled *push, lift, show*.

As shown in Fig. 11, the progress in recognition is better with the curiosity-driven exploration than with random exploration, for both teachers. At the end of the experiments, the SGIM-ACTS learner is able to correctly recognize the objects in 57 over 64 images, against 50 in the case of the random learner.

Not surprisingly, Fig. 12 shows that, when exploration is random, the object is changed more frequently, whereas when exploration is autonomous the robot focuses on objects for longer periods. In the "random" case the robot does not focus on any particular object: since it explores equally all objects, the recognition performance at the end of the experiment is worse, because the "difficult" objects (such as the cubes - green line) are not sufficiently explored. Conversely, the SGIM-ACTS learner focuses more on the difficult objects such as the cubes, especially when its competence progress increases. Fig. 12c and 12d clearly illustrate this mechanism: the red bear (cyan line) is easily recognized, hence the robot does not ask again to interact with the object once it is learnt; conversely, the cubes (green line) are difficult to recognize, since their appearance changes substantially depending on the action (a frontal view consists of four cubes, while a lateral view consists of two cubes only, and depending on the side it could be yellow or red/green), hence the robot focuses more on them. For both teachers, the robot spent 54% and 51% of its time learning about cubes when exploration was curiosity-driven. This proves that intrinsic motivation makes the robot focus on the most difficult objects to learn.

The curiosity mechanism is also necessary to compensate for good or bad teaching actions: this is a crucial point, because it allows the robot to take advantage of the coaching of experienced researchers but also collaborate with naive subjects. With the "good" teacher (unbiased) the robot decided to autonomously do 50.85% push, 23.73% take/lift/throw, and asked the human to do 25.42% manipulate/show. With the "bad" teacher (biased) the robot did autonomously 22.97% push, 40.54% take/lift/throw, and asked the human to do 36.49% manipulate/show.

Notably, with the "bad" teacher the robot takes and throws more the objects (41% vs 24%) to compensate with its active manipulation the lack of informative input from the teacher.

A "good" teacher can thus have a catalyzing effect: the learning process is 25% faster with an unbiased teacher than with the biased one, and the robot can focus on manipulating more the complex objects. But, thanks to the curiosity mechanism, the teaching component is not fundamental to determine the final outcome of the learning process: as shown in Fig. 11, the curiosity-driven exploration allows the robot to learn efficiently all the objects with both teachers.

[12] On average, one experiment takes between 60 and 90 minutes.

[13] Since speech recognition is not integrated in this experiment, we manually enter the labels of the objects, i.e. their names, into the curiosity module. This step is specific to this experiment, and is necessary to ensure that the robot knows that during this session only those five objects will be available. We remind that since the robot has limited mobility, it has to ask the human to show and put the objects on the table each time. To ease the communication with the human and simplify the experiment focusing on the exploration strategy, we chose to give the objects' names to the robot at the beginning of the experiment.

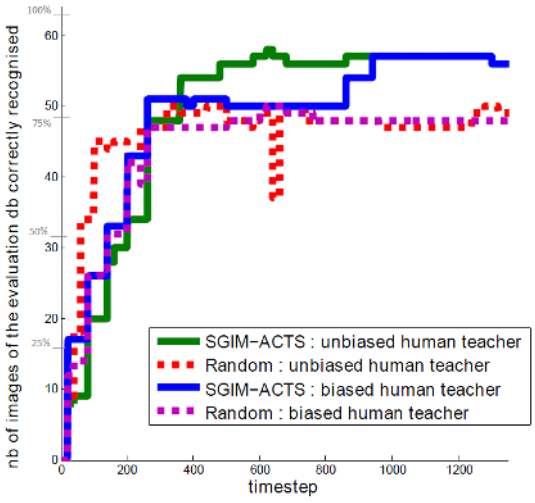


Fig. 11. SGIM-ACTS vs Random: recognition performance, i.e. the number of images of the evaluation database correctly recognized by the two exploration strategies with two different behaviors of the teacher (see text).

IV. DISCUSSION

The experimental results of Section III highlight two important observations that we discuss hereinafter.

A. Learning relates to manipulation

The influence of motor activity in learning about object representations and properties has been widely investigated in psychology, especially since Gibson's arguments that infants discover the affordances of the objects through motor activity [6], [73]. Oakes and Baumgartner [3] showed that the quality of infants' manipulation is related to the way they learn about object properties. Particularly, they showed that the more infants want to learn about an object, the more time they manipulate and touch the object. These observations find correspondences in our experiment: the more the robot wants to learn about an object, the more the object is complex or informative, the more time it decides to manipulate the object (see Section III-B). Moreover, the robot chose privileged "informative" actions yielding the more possible changes in the appearance of the object, like pushing or throwing, to cope with different types of social partners. This puts forward two interesting points. First, the combination of social guidance and intrinsic motivation is a promising approach for building robots capable of learning from and in cooperation with naive subjects (typically, non-researchers, non-roboticists who have no idea about how to leverage their learning process). Second, object's properties and affordances can emerge from the interaction of the robot with the object. The experiments of Section III were focused on objects recognition, so we looked at the object's representation before and after an action affects the status of the object. However, the observation of effects is a necessary step towards the learning of more complicated properties of the objects, such as affordances. Fig. 13 reports the median displacement of the objects pushed by the robot during the experiments of Section III-A. A simple classification of the effect defined as the $y-$axis displacement can provide information about the object's structure. In this case, the presence of a main axis in the object's shape can help the robot to generalize and transfer the acquired "rollable" property to other objects of similar shape. These ideas are at the base of research and experiments on the acquisition and exploitation of objects' affordances [74], [75], [76]; moreover, they could be easily integrated with symbol and language grounding [77]. Such experiments are out of the scope of this paper, but they are one of the natural follow-up to our work.

B. The observation of effects due to action in a spatial context

Section III-B presented our approach for the observation of the visual effect of actions performed on objects by a cognitive agent. This approach is grounded on the way the robot builds its perceptual representation of the objects, which was described in Section II-D. As shown in Fig. 3, the processing of perceptual information is based on a layered architecture, which extracts visual features from low-level sensory data (vision, proprioception) and elaborates them to determine the objects in the scene. Objects are then represented in a robot-centric reference frame, so that the robot can use their spatial information for planning motions and interacting with the objects. This robot-centric representation is convenient for manipulating the objects, however it may not be the most convenient to study the effect of the actions on the objects or the objects' affordances. As discussed in [78], when perception relies on static images, spatial context information is necessary to help the cognitive agents to observe and make inferences. For example, if the context is the one depicted in Fig. 14, the effect of the action is a relative displacement between the two objects. This being the goal of the action, reasoning should occur in a non-robot-centric reference frame but rather in a frame related to the context of the problem. In that case, the same cognitive process could be used not only to identify the effect of robot's action on the objects (which is easy because the robot knows which action it is doing and to which object), but also to infer which is the best action-object couple that produce the desired effect (the relative displacement). For analogous arguments about planning in the perceptual space using affordances we refer to [79] for a survey.

C. Multimodal hierarchical representations for cognition

The cognitive elements described in Section II are grounded on a multimodal hierarchical representation of the robot state and environment through its sensors [14]. As illustrated in Fig. 3, our vision processing system has a hierarchical structure in the way images are elaborated so as to provide the final information about the object, its identity, its position in the robot's frame etc. The introduction of motor information makes the spaces larger, but as shown in Section II-D, it enables enriching the object's representation with categories (*human body*, *manipulable object*, *robot body*), which are in this case the product of a sensorimotor coupling.

In literature, it is still debated which structures should be used to represent sensorimotor couplings and promote the

[14]The concept evokes the sensory ego-sphere [63], however our representation is structurally different.

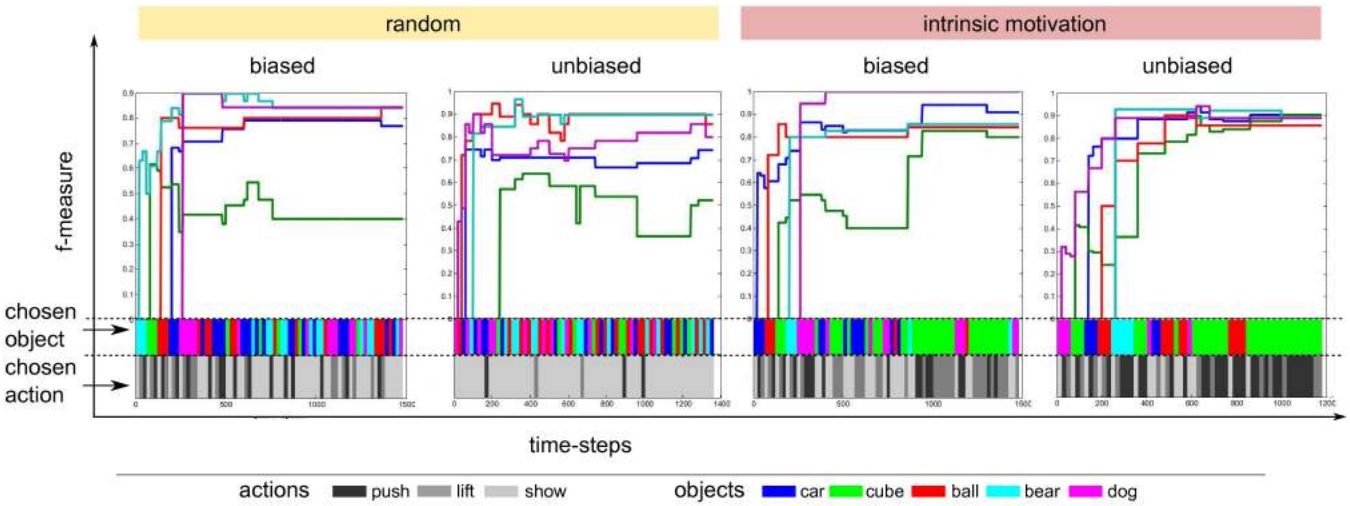


Fig. 12. A comparison between random (left) and SGIM-ACTS-based, curiosity-driven (right) exploration with biased and unbiased teacher. The evolution of the f-measure (see text) computed on the evaluation database is plotted with respect to time, during the course of one experiment per case. The bottom rows of each plot indicate with a color code the manipulated object and the chosen action to perform at each time-step.

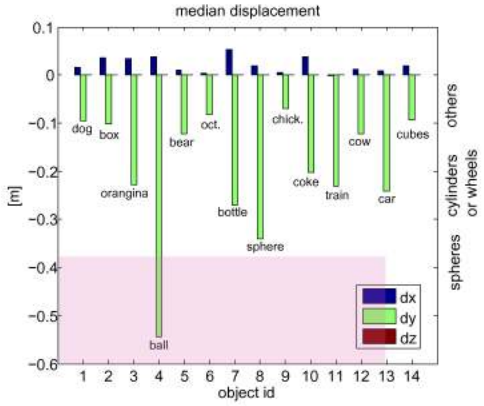


Fig. 13. The median displacement of the objects of Fig. 8, due to the *push-left* action (30 trials for each object). Two thresholds on the y-axis displacement, at 0.3m and 0.15m, reveal a rough categorization of the objects' "rolling" property. Objects that moved more than 0.3m "rolled" because they have not a preferred direction of movement: indeed, they are spherical. Objects that moved in between the two thresholds had at least one axis that was favorable for motion: indeed, these objects had either a cylindrical shape (hence they can roll in the direction normal to their main axis) or wheels (the rolling direction is perpendicular to the wheels). The other objects did not roll because their shape was unfit for this motion.

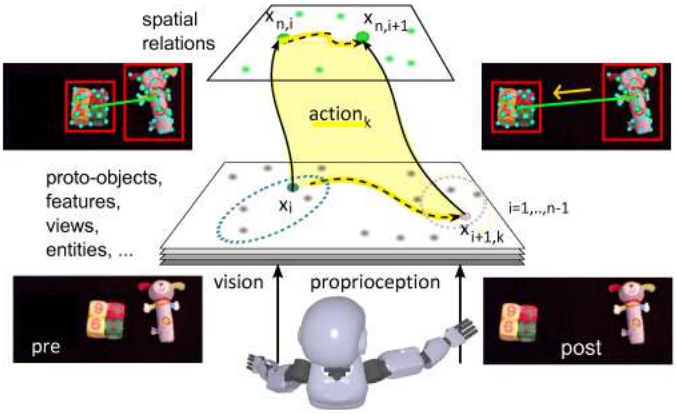


Fig. 14. The observation of the effects of actions in a spatial context. The visual processing pipeline (Fig. 3) used for the *push-left* experiment eventually represents entities in a robot-centric frame. In different contexts, where for example one object moves with respect to another, it is more convenient to represent the effect of an action in terms of relative spatial transformation between two (or more) entities in the visual scene.

emergence of a structure in the representation of the robot's body and environment [80]. Several approaches can be found. In the variety of emergentist approaches, one can find a blend of neural implementations, probabilistic structures and AI planning algorithms, applied to learning low-level sensorimotor functions [81], [82], causal relationships between objects and actions [74], [75], [83], [84], [85], [86], *etc*. All these techniques are valuable and capable of solving the problem at their low, middle or high level of abstraction from the sensorimotor interfaces of the robot. However, they often define a priori the interconnections between each computation layer (as we usually do between software modules or middlewares), which reduces the dimensionality of the problem. This has been done for example in [79] for the discovery of objects affordances, where the authors chose a set of variables for detecting the effects of actions on objects (e.g. position, shape, visibility). Quoting Montesano *et al.* in [74], "*learning from scratch can be overwhelming, as it involves relations between motor and perceptual skills, resulting in an extremely large dimension search problem.*" This issue can be counteracted if perceptual and learning processes are simplified and decoupled when possible. However, by letting the structure define how the development of each level influences the progress of the others, the robot can develop its own representation of states and transitions, in an intertwined way [87]. More importantly, the cognitive process can learn to build semantic representations of each of the robot's perceptual modality and fuse them in higher abstract levels of knowledge [29]. This opens the possibility of efficient unsupervised learning, where modalities provide "labels" to each other [88]. We believe our hierarchical and

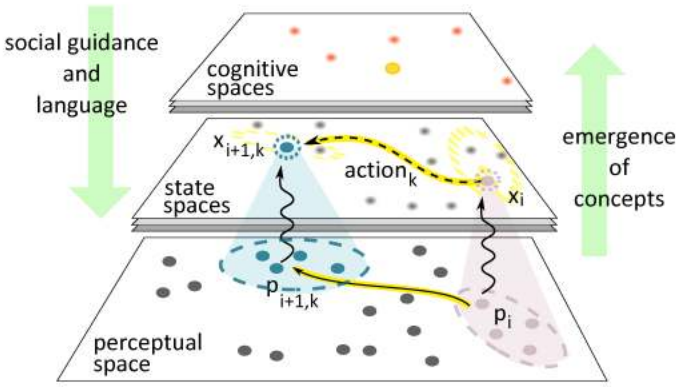


Fig. 15. Emergence of cognitive representations in a hierarchical architecture. In the picture, two *pre* and *post* states are identified by a perceptual representation p (i.e. the elaboration of the sensory inputs) and a state representation x. The action a_k acting on state p_i (or x_i) provokes an effect $e_{i,k}$, which can belong to a higher cognitive space. Navigating in the bottom-up direction the hierarchy, information is processed and becomes more and more abstract: perceptual data are elaborated (e.g. objects and their categories are extracted), gradually categories and properties of the objects emerge. The more and more abstracted concepts/symbols can be used to ground language. Indeed, social drives and language can be used to navigate top-bottom the hierarchy, to guide learning and exploration process, to trim the state spaces, etc.

multimodal representation can provide an interesting base for the emergence of knowledge in that way. This particular point is discussed in the paper's perspectives.

V. CONCLUSION AND FUTURE WORK

In this paper we presented the cognitive architecture of the MACSi project, which is designed to combine intrinsic motivation and social guidance with perceptual and motor learning. Our architecture brings a novel contribution in the way multimodal perception and motor commands are combined in a developmental fashion, where knowledge emerges in a hierarchical multimodal structure. We presented the architecture in Section II, describing its main functionalities.

We presented some experiments with the humanoid iCub, where we exploit the hallmarks of our architecture to make the robot learn autonomously about objects through active exploration. Section III-A showed why manipulation is important to comprehensively explore objects, while Section III-B showed how curiosity and social guidance can guide this exploration. Experimental results show the effectiveness of our approach: the humanoid iCub is capable of deciding autonomously which objects should be explored and which manipulating actions should be performed in order to improve its knowledge, requiring a minimal assistance from its caregiver. As discussed in Section IV, our results relate to observations in infants development.

While we demonstrated the efficiency of the architecture for an object learning task, the architecture could be easily used to tackle more complex cognitive tasks, for example learning objects' affordances and grounding symbols and language through active exploration and social interaction. As we discussed in Section IV-C, the multimodal hierarchical processing enables enriching the object's representation with categories. That is, the cognitive process can build semantic representations, where modalities provide "labels" to each other, in an unsupervised way.

One promising solution to evolve our cognitive structure is by means of deep hierarchical structures of multi-layer neural networks, which encounter an increasing popularity in the deep-learning community [89], [90], [91]. Their features are very interesting for a developmental learning process: intrinsic hierarchical structure and semantic abstractions [92], intertwined learning of states and transitions [93], multimodality [94], [95], and they are very weakly supervised [96]. Moreover, from a developmental perspective, a parallel can be found between the need for successive training of each layer of deep neural networks [97] and the emergence of refined outcomes at different timings in the development of the child [5], [98].

Remarkably, the use of hierarchical representations favors the grounding of symbols and language [41]. As explained in [99], it is debated whether the ability of infants to acquire language relies on their ability to build recursive structures or on their progress in the knowledge of items and abstractions. For a cognitive agent, a combination of both is likely necessary. As illustrated in Fig. 15, language can provide a top-down signal in the hierarchy of perceptual and cognitive states that can help refining the learned representations at the lower levels and defining multimodal mappings. It naturally provides a "label" to the unsupervised learning process: for example, in [100], abstract symbols are used at the highest level of a deep neural network to provide one shot learning of new categories of objects. More generally, such representations could be ideally integrated in high-level shared databases and frameworks, such as RoboEarth [101], [102].

In summary, we envisage two main improvements. First, the evolution of the architecture towards a more complex hierarchical structure, which exploits the latest results in deep learning community [89], to classify objects and binding language with visuo-motor representations. Second, the combination of physical and social guidance with language in a developmental fashion [41] to reproduce what has been called "parental scaffolding" [16].

VI. CODE, DOCUMENTATION AND VIDEOS

The software implementing the cognitive architecture and all the experiments is open-source, available under GPL license at http://macsi.isir.upmc.fr. Instruction for running the code and replicate the experiments can be found at http://chronos.isir.upmc.fr/~ivaldi/macsi/doc/. Videos can be downloaded at http://www.youtube.com/user/iCubParis/videos.

APPENDIX A
SGIM-ACTS

We hereby describe how the intrinsic motivation system, outlined in Section II-B, is used to find the exploration strategy *(object, action, actor)* for the object learning experiment described in Section III-B.

Let us consider images $a \in A$, objects $b \in B$, and the mapping $M : A \mapsto B$ defining the correct labelling of all images with the objects' names. For each image perceived by

the camera, the iCub computes the likelihood for each already known views, and returns the two highest likelihood measures p_1, p_2, as well as the labels b_1, b_2 of the objects associated with the views, and the number n_1, n_2 of known views for each of the labels. The label b_g of the currently manipulated object is known to robot, as it is tough by the teacher in a preliminary phase. The robot estimates its competence at distinguishing b_g from other objects, with the dissimilarity of likelihood measures between the 1st object associated and the 2nd object associated, and by estimating its gain of information about the object by collecting new views. The competence is defined as

$$\begin{aligned}\gamma(b_g) =& n_1 \times p_1 + c_1 && \text{if } b_g = b_1 = b_2\\ & n_1 \times p_1/(1+p_2) + c_1 && \text{if } b_g = b_1, b_g \neq b_2\\ & n_2 \times p_2/(1+p_1) + c_1 && \text{if } b_g \neq b_1, b_g = b_2\\ & c_1 && \text{if } b_g \neq b_1, b_g \neq b_2\end{aligned}$$

where c_1 is a constant, set to -1 in our experiment.

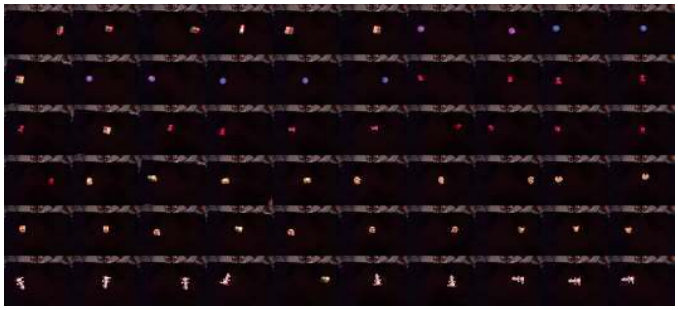
Our learner improves the estimation L of M to maximize $I = \sum_a P(a)\gamma(a)$, both by self-exploring A and B spaces by generating new perception samples through manipulation of the objects and by asking for help to a caregiver, who handles the objects to the robot. When an object is placed on the table, an image $a \in A$ of an object $b \in B$ is retrieved at each step. SGIM-ACTS learns by episodes during which it actively chooses both an object $b \in B$ to learn to recognize and a learning strategy σ between: pushing the object, taking and dropping the object or asking the caregiver to manipulate the object. For each object b it has decided to explore, it also decides the strategy σ which maximizes its "competence progress" or "interest", defined as the local competence progress, over a sliding time window of δ for an object b with strategy σ at cost $\kappa(\sigma)$. If the competence measured for object b with strategy σ constitute the list $R(b,\sigma) = \{\gamma_1, ...\gamma_N\}$:

$$interest(b,\sigma) = \frac{1}{\kappa(\sigma)} \frac{\left| \left(\sum_{j=N-\delta}^{N-\frac{\delta}{2}} \gamma_j \right) - \left(\sum_{j=N-\frac{\delta}{2}}^{N} \gamma_j \right) \right|}{\delta}$$

This strategy enables the learner to generate new samples a in subspaces of A. The SGIM-ACTS learner explores preferentially objects where it makes progress the fastest. More details of the algorithm and its implementation can be found in [25].

APPENDIX B
DATABASE FOR EVALUATION OF CURIOSITY PERFORMANCES

We evaluate the progress in objects' recognition by computing a performance measure over an evaluation database (Fig. 16). It must be noticed that the evaluation of the learning progress on the database is not used directly by the curiosity system to guide the exploration (see details in [103]), but is mostly used to visualize the learning curves of the experiments. The database consists of 64 images acquired by the RGB-D camera before the learning process takes place. In this preparation phase, the human caregiver showed all the five objects (multi-colored cubes, blue-violet sphere, red bear, gray dog, yellow car) under several and different views (Fig. 16b). It must be remarked that the whole image taken by the RGB-D camera is retained for the evaluation (Fig. 16a).



(a) Evaluation database - (60 out of 64 images)

(b) Aggregation of the objects' views in the evaluation database

Fig. 16. The database of objects' views used by the intrinsic motivation system to evaluate the progress in object recognition. It consists of 64 random images of the objects from different points of view (a), which are conveniently regrouped in (b).

REFERENCES

[1] D. Vernon, C. von Hofsten, and L. Fadiga, *A roadmap for cogniive development in humanoid robots*. Springer, 2010.

[2] P. H. Miller, *Theories of developmental psychology*, 5th ed. Worth Publishers, 2010.

[3] L. M. Oakes and H. A. Baumgartner, "Manual object exploration and learning about object features in human infants," in *IEEE International Conference on Development and Learning and Epigenetic Robotics, ICDL-EPIROB*, 2012, pp. 1–6.

[4] E. Bushnell and R. Boudreau, "Motor development and the mind: the potential role of motor abilities as a determinant of aspects of perceptual development," *Child Development*, vol. 64(4), pp. 1005–1021, 1993.

[5] F. Guerin, N. Krueger, and D. Kraft, "A survey of the ontogeny of tool use: from sensorimotor experience to planning," *IEEE Transactions on Autonomous Mental Development*, vol. 5(1), pp. 18–45, 2013.

[6] E. J. Gibson, "Exploratory Behavior In The Development Of Perceiving, Acting, And The Acquiring Of Knowledge," *Annual Review of Psychology*, vol. 39, pp. 1–41, 1988.

[7] K. Bourgeois, A. Khawar, S. Neal, and J. Lockman, "Infant manual exploration of objects, surfaces, adnd their interrelations," *Infancy*, vol. 8(3), pp. 233–252, 2005.

[8] R. Wu, A. Gopnik, D. Richardson, and N. Kirkham, "Infants learn about objects from statistics and people," *Developmental Psychology*, vol. 47(5), pp. 1220–1229, 2011.

[9] P. Fitzpatrick, G. Metta, L. Natale, S. Rao, and G. Sandini, "Learning about objects through action - initial steps towards artificial cognition," in *IEEE Int. Conf. on Robotics and Automation*, 2003, pp. 3140–3145.

[10] M. Lungarella and O. Sporns, "Mapping information flow in sensorimotor networks," *PLoS Comput Biol*, vol. 2, no. 10, p. e144, 10 2006.

[11] M. Lopes and J. Santos-Victor, "Learning sensory-motor maps for redundant robots," in *IROS*. IEEE, 2006, pp. 2670–2676.

[12] N. Lyubova and D. Filliat, "Developmental approach for interactive object discovery," in *Int. Joint Conf. on Neural Networks*, 2012.

[13] M. Lopes and P. Oudeyer, "Active learning and intrinsically motivated exploration in robots: Advances and challenges (guest editorial)," *IEEE Trans. on Autonomous Mental Development*, vol. 2, no. 2, pp. 65–69, 2010.

[14] J. Schmidhuber, "Formal theory of creativity, fun, and intrinsic motivation (1990-2010)," *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 3, pp. 230–247, 2010.

[15] A. L. Thomaz, "Socially guided machine learning," Ph.D. dissertation, MIT, 5 2006.

[16] E. Ugur, H. Celikkanat, E. Sahin, Y. Y. Nagai, and E. Oztop, "Learning to grasp with parental scaffolding," in *IEEE Int. Conf. on Humanoid Robotics*, 2011, pp. 480–486.

[17] D. Vernon, G. Metta, and G. Sandini, "A survey of artificial cognitive systems: Implications for the autonomous development of mental capabilities in computational agents," *IEEE Trans. Evolutionary Computation*, vol. 11(2), pp. 151–180, 2007.

[18] M. Malfaz, A. Castro-Gonzalez, R. Barber, and M. Salichs, "A biologically inspired architecture for an autonomous and social robot," *IEEE Trans. Aut. Mental Development*, vol. 3, no. 3, pp. 232 –246, 2011.

[19] S. Lallée, U. Pattacini, J.-D. Boucher, S. Lemaignan, A. Lenz, C. Melhuish, L. Natale, S. Skachek, K. Hamann, J. Steinwender, E. A. Sisbot, G. Metta, R. Alami, M. Warnier, J. Guitton, F. Warneken, and P. F. Dominey, "Towards a platform-independent cooperative human-robot interaction system: Ii. perception, execution and imitation of goal directed actions," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2011, pp. 2895–2902.

[20] S. Rockel, B. Neumann, J. Zhang, K. S. R. Dubba, A. G. Cohn, S. Konecny, M. Mansouri, F. Pecora, A. Saffiotti, M. Gunther, S. Stock, J. Hertzberg, A. M. Tome, A. Pinho, L. Seabra Lopes, S. von Riegen, and L. Hotz, "An ontology based multi-level robot architecture for learning from experiences," in *AAAI Spring Symposium 2013 on Designing Intelligent Robots: Reintegrating AI*, 2013.

[21] S. Lemaignan, R. Ros, L. Mosenlechner, R. Alami, and M. Beetz, "Oro, a knowledge management module for cognitive architectures in robotics," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010.

[22] S. Ivaldi, N. Lyubova, D. Gérardeaux-Viret, A. Droniou, S. M. Anzalone, M. Chetouani, D. Filliat, and O. Sigaud, "Perception and human interaction for developmental learning of objects and affordances," in *Proc. IEEE-RAS Int. Conf. on Humanoid Robots*, Osaka, Japan, 2012.

[23] A. Droniou, S. Ivaldi, V. Padois, and O. Sigaud, "Autonomous online learning of velocity kinematics on the icub: a comparative study," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems - IROS*, 2012, pp. 5377–5382.

[24] S. M. Anzalone, S. Ivaldi, O. Sigaud, and M. Chetouani, "Multimodal people engagement with iCub," in *Proc. Int. Conf. on Biologically Inspired Cognitive Architectures*, Palermo, Italy, 2012.

[25] S. M. Nguyen and P.-Y. Oudeyer, "Active choice of teachers, learning strategies and goals for a socially guided intrinsic motivation learner," *Paladyn Journal of Behavioural Robotics*, vol. 3, no. 3, pp. 136–146, 2012.

[26] J. Weng, J. McClelland, A. Pentland, O. Sporns, I. Stockman, M. Sur, and E. Thelen, "Autonomous mental development by robots and animals," *Science*, vol. 291, pp. 599–600, 2001.

[27] P.-Y. Oudeyer, F. Kaplan, and V. Hafner, "Intrinsic motivation systems for autonomous mental development," *IEEE Transactions on Evolutionary Computation*, vol. 11(2), pp. pp. 265–286, 2007.

[28] J. Piaget, *The Origins of Intelligence in Children*. New York: International University Press, 1952.

[29] L. Smith and M. Gasser, "The development of embodied cognition: six lessons from babies," *Artificial Life*, vol. 11, pp. 13–29, 2005.

[30] L. Natale, F. Nori, G. Metta, M. Fumagalli, S. Ivaldi, U. Pattacini, M. Randazzo, A. Schmitz, and G. G. Sandini, *Intrinsically motivated learning in natural and artificial systems*. Springer-Verlag, 2012, ch. The iCub platform: a tool for studying intrinsically motivated learning.

[31] G. Bremner and A. Fogel, Eds., *Handbook of Infant Development*. Blackwell, 2001.

[32] J. Konczak and J. Dichgans, "The development toward stereotypic arm kinematics during reaching in the first 3 years of life," *Experimental Brain Research*, vol. 117, pp. 346–354, 1997.

[33] N. Lyubova, S. Ivaldi, and D. Filliat, "Developmental object learning through manipulation and human demonstration," in *Interactive Perception Workshop - IEEE/RAS Int. Conf. on Robotics and Automation*, 2013, p. 1.

[34] H. Ruff, L. Saltarelli, M. Capozzoli, and K. Dubiner, "The differentiation of activity in infants' exploration of objects," *Developmental psychology*, vol. 28(5), pp. 851–861, 1992.

[35] S. Ivaldi, M. Fumagalli, M. Randazzo, F. Nori, G. Metta, and G. Sandini, "Computing robot internal/external wrenches by means of inertial, tactile and F/T sensors: theory and implementation on the iCub," in *Proc. of IEEE-RAS Int. Conf. on Humanoid Robots*, 2011, pp. 521–528.

[36] M. Fumagalli, S. Ivaldi, M. Randazzo, L. Natale, G. Metta, G. Sandini, and F. Nori, "Force feedback exploiting tactile and proximal force/torque sensing. theory and implementation on the humanoid robot icub," *Autonomous Robots*, vol. 33, no. 4, pp. 381–398, 2012.

[37] B. Berret, S. Ivaldi, F. Nori, and G. Sandini, "Stochastic optimal control with variable impedance manipulators in presence of uncertainties and delayed feedback," in *Proc. of the 2011 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems - IROS*, San Francisco, CA, USA, 2011, pp. 4354–4359.

[38] A. Cully and J.-B. Mouret, "Behavioral repertoire learning in robotics," in *Proc. Int. Conf. on Genetic and Evolutionary Computation*, 2013.

[39] J. Lehman and K. Stanley, "Abandoning objectives: evolution through the search for novelty alone," *Evolutionary computation*, vol. 19(2), pp. 189–223, 2011.

[40] A. Edsinger and C. Kemp, "What can i control? a framework for robot self-discovery," in *Proc. Int. Conf. Epigenetic Robotics*, 2006.

[41] A. Cangelosi, G. Metta, G. Sagerer, S. Nolfi, C. Nehaniv, K. Fischer, J. Tani, T. Belpaeme, G. Sandini, F. Nori, *et al.*, "Integration of action and language knowledge: A roadmap for developmental robotics," *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 3, pp. 167–195, 2010.

[42] F. Stulp, G. Raiola, A. Hoarau, S. Ivaldi, and O. Sigaud, "Learning compact parameterized skills with expanded function approximators," in *Proc. IEEE-RAS Int. Conf. on Humanoid Robots*, 2013.

[43] E. Deci and R. M. Ryan, *Intrinsic Motivation and self-determination in human behavior*. New York: Plenum Press, 1985.

[44] R. M. Ryan and E. L. Deci, "Intrinsic and extrinsic motivations: Classic definitions and new directions," *Contemporary Educational Psychology*, vol. 25, no. 1, pp. 54 – 67, 2000.

[45] J. Schmidhuber, "Curious model-building control systems," in *Proc. Int. Joint Conf. Neural Netw.*, vol. 2, 1991, pp. 1458–1463.

[46] V. Fedorov, *Theory of Optimal Experiment*. New York, NY: Academic Press, Inc., 1972.

[47] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," *Journal of Artificial Intelligence Research*, vol. 4, pp. 129–145, 1996.

[48] N. Roy and A. McCallum, "Towards optimal active learning through sampling estimation of error reduction," in *Proc. 18th Int. Conf. Mach. Learn.*, vol. 1, 2001, pp. 143–160.

[49] C. Rich, A. Holroyd, B. Ponsler, and C. Sidner, "Recognizing engagement in human-robot interaction," in *ACM/IEEE International Conference on Human Robot Interaction*, 2010, pp. 375–382.

[50] P.-Y. Oudeyer and F. Kaplan, "What is intrinsic motivation? a typology of computational approaches," *Frontiers of Neurorobotics*, p. 1:6, 2007.

[51] M. Lopes and P.-Y. Oudeyer, "The Strategic Student Approach for Life-Long Exploration and Learning," in *IEEE Conference on Development and Learning / EpiRob*, San Diego, États-Unis, Nov. 2012.

[52] D. Forte, A. Gams, J. Morimoto, and A. Ude, "On-line motion synthesis and adaptation using a trajectory database," *Robotics and Autonomous Systems*, vol. 60, pp. 1327–1339, 2012.

[53] M. Petit, S. Lallee, J.-D. Boucher, G. Pointeau, P. Cheminade, D. Ognibene, E. Chinellato, U. Pattacini, I. Gori, U. Martinez-Hernandez, H. Barron-Gonzalez, M. Inderbitzin, A. Luvizotto, V. Vouloutsi, Y. Demiris, G. Metta, and P. Dominey, "The coordinating role of language in real-time multimodal learning of cooperative tasks," *Autonomous Mental Development, IEEE Transactions on*, vol. 5, no. 1, pp. 3–17, 2013.

[54] N. Lyubova, D. Filliat, and S. Ivaldi, "Improving object learning through manipulation and self-identification," in *Submitted*, 2013.

[55] Z. W. Pylyshyn, "Visual indexes, preconceptual objects, and situated vision," *Cognition*, vol. 80, pp. 127–158, 2001.

[56] J. Shi and C. Tomasi, "Good features to track," in *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on*, 1994, pp. 593–600.

[57] H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," in *In ECCV*, 2006, pp. 404–417.

[58] B. Micusik and J. Kosecka, "Semantic segmentation of street scenes by superpixel co-occurrence and 3d geometry," in *IEEE Int. Conf. on Computer Vision*, 2009, pp. 625–632.

[59] D. Filliat, "A visual bag of words method for interactive qualitative localization and mapping," in *IEEE Int. Conf. on Robotics and Automation*, 2007, pp. 3921–3926.

[60] J. Sivic and A. Zisserman, "Video google: Text retrieval approach to object matching in videos," in *Int. Conf. on Computer Vision*, vol. 2, 2003, pp. 1470–1477.

[61] C. Kemp and A. Edsinger, "What can i control?: The development of visual categories for a robots body and the world that it influences," in *IEEE Int. Conf. on Development and Learning (ICDL), Special Session on Autonomous Mental Development*, 2006.

[62] S. Lallee, U. Pattacini, S. Lemaignan, A. Lenz, C. Melhuish, L. Natale, S. Skachek, K. Hamann, J. Steinwender, E. Sisbot, G. Metta, J. Guitton, R. Alami, M. Warnier, T. Pipe, F. Warneken, and P. Dominey, "Towards a platform-independent cooperative human robot interaction system: Iii an architecture for learning and executing actions and shared plans," *IEEE Transactions on Autonomous Mental Development*, vol. 4, no. 3, pp. 239–253, sept. 2012.

[63] R. A. Peters, Ii, K. A. Hambuchen, and R. E. Bodenheimer, "The sensory ego-sphere: a mediating interface between sensors and cognition," *Auton. Robots*, vol. 26, no. 1, pp. 1–19, 2009.

[64] J. Ruesch, M. Lopes, A. Bernardino, J. Hornstein, J. Santos-Victor, and R. Pfeifer, "Multimodal saliency-based bottom-up attention a framework for the humanoid robot icub," in *IEEE Int. Conf. Robotics and Automation*, 2008, pp. 962–967.

[65] U. Pattacini, F. Nori, L. Natale, G. Metta, and G. Sandini, "An experimental evaluation of a novel minimum-jerk cartesian controller for humanoid robots." in *Proc. of IEEE/RSJ Int. Conf. Intelligents Robots and Systems*, 2010, pp. 1668–1674.

[66] P. Fitzpatrick, G. Metta, and L. Natale, "Towards long-lived robot genes," *Robotics and Autonomous Systems*, vol. 56, no. 1, pp. 29–45, 2008.

[67] www.ros.org.

[68] S. Perone, K. Madole, S. Ross-Sheehy, M. Carey, and L. M. Oakes, "The relation between infants' activity with objects and attention to object appearance," *Developmental Psychology*, vol. 44, pp. 1242–1248, 2008.

[69] A. L. Thomaz and M. Cakmak, "Learning about objects with human teachers," in *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. ACM, 2009, pp. 15–22.

[70] C. Breazeal and A. L. Thomaz, "Learning from human teachers with socially guided exploration," in *IEEE International Conference on Robotics and Automation*. IEEE, 2008, pp. 3539–3544.

[71] S. M. Nguyen, A. Baranes, and P.-Y. Oudeyer, "Bootstrapping intrinsically motivated learning with human demonstrations," in *IEEE Int. Conf. on Development and Learning*, 2011.

[72] C. J. van Rijsbergen, *Information Retrieval (2nd ed.)*. Butterworth, 1979.

[73] J. Gibson, *Perceiving, Acting, and Knowing: Toward an Ecological Psychology (R. Shaw & J. Bransford Eds.)*. Lawrence Erlbaum, 1977, ch. The Theory of Affordances, pp. 67–82.

[74] L. Montesano, M. Lopes, A. Bernardino, and J. Santos-Victor, "Learning object affordances: From sensory–motor coordination to imitation," *IEEE Transactions on Robotics*, vol. 24, no. 1, pp. 15 –26, feb. 2008.

[75] E. Ugur and E. Sahin, "Traversability: A case study for learning and perceiving affordances in robots," *Adaptive Behavior*, vol. 18, pp. 258–284, 2010.

[76] C. Castellini, T. Tommasi, N. Noceti, F. Odone, and B. Caputo, "Using object affordances to improve object recognition," *IEEE Trans. on Autonomous Mental Development*, vol. 3, no. 3, pp. 207 –215, 2011.

[77] O. Yuruten, K. F. Uyanik, Y. Caliskan, A. K. Bozcuoglu, E. Sahin, and S. Kalkan, "Learning adjectives and nouns from affordances on the icub humanoid robot," in *From Animals to Animats*, ser. Lecture Notes in Computer Science, T. Ziemke, C. Balkenius, and J. Hallam, Eds., vol. 7426. Springer Berlin Heidelberg, 2012, pp. 330–340.

[78] E. Zibetti and C. Tijus, "Perceiving action from static images: The role of spatial context," in *Modeling and Using Context*, ser. Lecture Notes in Computer Science, P. Blackburn, C. Ghidini, R. Turner, and F. Giunchiglia, Eds. Springer Berlin Heidelberg, 2003, vol. 2680, pp. 397–410.

[79] E. Ugur, E. Oztop, and E. Sahin, "Goal emulation and planning in perceptual space using learned affordances," *Robotics and Autonomous Systems*, vol. 59, no. 7-8, pp. 580–595, 2011.

[80] J. Tenenbaum, C. Kemp, T. Griffiths, and N. Goodman, "How to grow a mind: statistics, structure and abstraction," *Science*, vol. 331, pp. 1279–1285, 2011.

[81] T. Waegeman and B. Schrauwen, "Towards learning inverse kinematics with a neural network based tracking controller," in *Neural Information Processing*, ser. Lecture Notes in Computer Science, B.-L. Lu, L. Zhang, and J. Kwok, Eds. Springer Berlin Heidelberg, 2011, vol. 7064, pp. 441–448.

[82] R. F. Reinhart and J. J. Steil, "Reaching movement generation with a recurrent neural network based on learning inverse kinematics for the humanoid robot icub," in *Int. Conf. on Humanoid Robots - Humanoids*, 2009.

[83] B. Moldovan, P. Moreno, M. van Otterlo, J. Santos-Victor, and L. De Raedt, "Learning relational affordance models for robots in multi-object manipulation tasks," in *IEEE International Conference on Robotics and Automation - ICRA*, may 2012, pp. 4373–4378.

[84] N. Krüger, C. Geib, J. Piater, R. Petrick, M. Steedman, F. Wörgötter, A. Ude, T. Asfour, D. Kraft, D. Omrčen, A. Agostini, and R. Dillmann, "Object-action complexes: Grounded abstractions of sensory-motor processes," *Robotics and Autonomous Systems*, vol. 59, no. 10, pp. 740–757, 2011, doi:10.1016/j.robot.2011.05.009.

[85] K. Mourao, L. S. Zettlemoyer, R. P. A. Petrick, and M. Steedman, "Learning strips operators from noisy and incomplete observations," in *Conf. on Uncertainty in Artificial Intelligence - UAI*, 2012, pp. 614–623.

[86] D. Kraft, N. Pugeault, E. Başeski, M. Popović, D. Kragić, S. Kalkan, F. Wörgötter, and N. Krüger, "Birth of the object: detection of objectness and extraction of object shape through object–action complexes," *International Journal of Humanoid Robotics*, vol. 5, no. 02, pp. 247–265, 2008.

[87] A. C. Schapiro, T. T. Rogers, N. I. Cordova, N. B. Turk-Browne, and M. M. Botvinick, "Neural representations of events arise from temporal community structure," *Nature Neuroscience*, vol. 16, no. 4, pp. 486–492, 2013.

[88] S. Lallee and P. F. Dominey, "Multi-modal convergence maps: from body schema and self-representation to mental imagery," *Adaptive Behavior*, 2013.

[89] G. E. Hinton, "Learning multiple layers of representation," *Trends in Cognitive Sciences*, vol. 11, pp. 428–434, 2007.

[90] Y. Bengio, "Learning deep architectures for AI," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009, also published as a book. Now Publishers, 2009.

[91] I. Arel, D. C. Rose, and T. P. Karnowski, "Research frontier: deep machine learning–a new frontier in artificial intelligence research," *Comp. Intell. Mag.*, vol. 5, no. 4, pp. 13–18, Nov. 2010. [Online]. Available: http://dx.doi.org/10.1109/MCI.2010.938364

[92] Y. Lee, S. J. Lee, and Z. Popović, "Compact Character Controllers," *ACM Transactions on Graphics*, vol. 28, no. 5, 2009.

[93] G. W. Taylor, G. E. Hinton, and S. T. Roweis, "Two Distributed-State Models For Generating High-Dimensional Time Series," *J. Mach. Learn. Res.*, vol. 12, pp. 1025–1068, 2011. [Online]. Available: http://dl.acm.org/citation.cfm?id=1953048.2021035

[94] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal Deep Learning," in *International Conference on Machine Learning (ICML)*, Bellevue, USA, 2011.

[95] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Neural Information Processing Systems (NIPS)*, 2012.

[96] R. Salakhutdinov, "Learning deep generative models," Ph.D. dissertation, Toronto, Ont., Canada, Canada, 2009, aAINR61080.

[97] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. Platt, and T. Hoffman, Eds. Cambridge, MA: MIT Press, 2007, pp. 153–160.

[98] M. Asada, K. Hosoda, Y. Kuniyoshi, H. Ishiguro, T. Inui, Y. Yoshikawa, M. Ogino, and C. Yoshida, "Cognitive developmental robotics: A survey," *IEEE Trans. Autonomous Mental Development*, vol. 1, no. 1, pp. 12–34, 2009.

[99] C. Bannard, E. Lieven, and M. Tomasello, "Modeling children's early grammatical knowledge," *Proceedings of the National Academy of Sciences*, vol. 106, pp. 17 284–17 289, 2009.

[100] R. R. Salakhutdinov, J. Tenenbaum, and A. Torralba, "Learning to Learn with Compound HD Models," in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. S. Zemel, P. Bartlett, F. C. N. Pereira, and K. Q. Weinberger, Eds., 2011, pp. 2061–2069.

[101] M. Tenorth, A. C. Perzylo, R. Lafrenz, and M. Beetz, "Representation and exchange of knowledge about actions, objects, and environments in

the roboearth framework," *IEEE Transactions on Automation Science and Engineering*, vol. 10, no. 3, pp. 643–651, 2013.

[102] M. Tenorth and M. Beetz, "Knowrob – a knowledge processing infrastructure for cognition-enabled robots. part 1: The knowrob system," *International Journal of Robotics Research*, vol. 32, no. 5, pp. 566 – 590, 2013.

[103] S. Nguyen, S. Ivaldi, N. Lyubova, A. Droniou, D. Gerardeaux-Viret, D. Filliat, V. Padois, O. Sigaud, and P.-Y. Oudeyer, "Learning to recognize objects through curiosity-driven manipulation with the icub humanoid robot," in *Proc. IEEE Conf. Development and Learning and Epigenetic Robotics*, 2013.

Serena Ivaldi is a postdoctoral researcher in University Pierre et Marie Curie, where she coordinates the experiments of MACSi and CODYCO Projects on iCub. She received the B.S. and M.S. degree in Computer Engineering, both with highest honors, at the University of Genoa (Italy) and her PhD in Humanoid Technologies in 2011, jointly at the University of Genoa and Italian Institute of Technology. There she also held a research fellowship in the Robotics, Brain and Cognitive Sciences Department. Her research is centered on humanoid robots interacting physically with humans and environment. She is interested in developmental robotics, blending learning, perception and motor control. Web: http://chronos.isir.upmc.fr/~ivaldi

Sao Mai Nguyen is currently a PhD student in the Flowers Team where she studies how to combine curiosity-driven exploration and socially guided exploration to learn various motor skills, focusing on active learning for a strategic student learner to learn several tasks with several strategies, and interactive learning to choose what, who and when to learn from social interaction. She holds a master degree in computer science from Ecole Polytechnique and a master degree in adaptive machine systems from Osaka University. She started research in cognitive developmental robotics in Asada Lab, Japan, focusing on the theory of mind. Her main topics of interest are developmental robotics, cognitive science, machine learning, imitation learning and intrinsic motivation. Web: http://nguyensmai.free.fr

Natalia Lyubova graduated from the Northern Arctic Federal University, Russia, in 2008, and obtained M.S. degree at the University of Eastern Finland in 2010. She is currently working on her PhD on visual perception for humanoid robots at Ecole Nationale Supérieure de Techniques Avancées-ParisTech. Her main research interests are computer vision, cognitive robotics, and developmental learning.

Alain Droniou graduated from the Ecole Polytechnique, France, in 2011. He is currently a PhD student in developmental robotics at the Institut des Systemes Intelligents et de Robotique in the University Pierre et Marie Curie in Paris. His research interests include humanoid robotics, developmental learning and deep-learning networks.

Vincent Padois is an associate professor of Robotics and Computer Science and a member of the Institut des Systèmes Intelligents et de Robotique (ISIR, UMR CNRS 7222) at Université Pierre et Marie Curie (UPMC) in Paris, France. In 2001, he receives both an engineering degree from the Ecole Nationale dIngnieurs de Tarbes (ENIT), France and his masters degree in Automatic Control from the Institut National Polytechnique de Toulouse (INPT), France. From 2001 to 2005, he is a PhD student in Robotics of the ENIT/INPT Laboratoire Gnie de Production. In 2006 and 2007, he is a post-doctoral fellow in the Stanford Artificial Intelligence Laboratory and more specifically in the group of Professor O. Khatib. Since 2007, his research activities at ISIR are mainly focused on the automatic design, the modelling and the control of redundant and complex systems such as wheeled mobile manipulators, humanoid robots as well as standard manipulators evolving under constraints in complex environments. He is also involved in research activities that aim at bridging the gap between adaptation and decision making techniques provided by Artificial Intelligence and low-level, reactive control. Since 2011, he holds the "Intervention Robotics" RTE/UPMC chair position.

David Filliat graduated from the Ecole Polytechnique in 1997 and obtained a PhD on bio-inspired robotics navigation from Paris VI university in 2001. After 4 years as an expert for the robotic programs in the French armament procurement agency, he is now professor at Ecole Nationale Suprieure de Techniques Avances ParisTech. Head of the Robotics and Computer Vision team, he obtained the Habilitation Diriger des Recherches en 2011. His main research interest are perception, navigation and learning in the frame of the developmental approach for humanoid and mobile robotics. http://www.ensta-paristech.fr/~filliat/

Pierre-Yves Oudeyer is Research Director at Inria and head of the Inria and Ensta-ParisTech FLOWERS team (France). Before, he has been a permanent researcher in Sony Computer Science Laboratory for 8 years (1999-2007). After working on computational models of language evolution, he is now working 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 is laureate of the ERC Starting Grant EXPLORERS. He is editor of the IEEE CIS Newsletter on Autonomous Mental Development, and associate editor of IEEE Transactions on Autonomous Mental Development, Frontiers in Neurorobotics, and of the International Journal of Social Robotics. Web: http://www.pyoudeyer.com and http://flowers.inria.fr.

Olivier Sigaud is professor in Computer Science at University Pierre et Marie Curie, Paris. His research interests include machine learning, human motor control and developmental robotics. He is centrally interested in the use of machine learning methods to endow a humanoid robot with the motor learning capabilities of humans. Olivier Sigaud is the prime investigator of the MACSi project (see http://macsi.isir.upmc.fr/) and one of the organizers of a French working group dedicated to "Robotics and Neuroscience" within the French robotics community. He is the author of over 100 referenced technical publications in computer science and robotics. He also holds a PhD in philosophy since 2004. For more information, see http://people.isir.upmc.fr/sigaud.