Improving object learning through manipulation and robot self-identification

Natalia Lyubova, David Filliat, Serena Ivaldi

▶ To cite this version:

Natalia Lyubova, David Filliat, Serena Ivaldi. Improving object learning through manipulation and robot self-identification. IEEE International Conference on Robotics and Biomimetics (ROBIO), Dec 2013, Shenzen, China. 10.1109/ROBIO.2013.6739655 . hal-00919649

HAL Id: hal-00919649
https://hal.science/hal-00919649

Submitted on 17 Dec 2013

Improving object learning through manipulation and robot self-identification

Natalia Lyubova[1], David Filliat[1], Serena Ivaldi[2]

Abstract— **We present a developmental approach that allows a humanoid robot to continuously and incrementally learn entities through interaction with a human partner in a first stage before categorizing these entities into objects, humans or robot parts and using this knowledge to improve objects models by manipulation in a second stage. This approach does not require prior knowledge about the appearance of the robot, the human or the objects. The proposed perceptual system segments the visual space into proto-objects, analyses their appearance, and associates them with physical entities. Entities are then classified based on the mutual information with proprioception and on motion statistics. The ability to discriminate between the robot's parts and a manipulated object then allows to update the object model with newly observed object views during manipulation. We evaluate our system on an iCub robot, showing the independence of the self-identification method on the robot's hands appearances by wearing different colored gloves. The interactive object learning using self-identification shows an improvement in the objects recognition accuracy with respect to learning through observation only.**

Key-words: developmental robotics, incremental learning, robot self-identification, interactive object exploration

I. INTRODUCTION

Future service robots will need the ability to work in different human environments that cannot be predicted in advance. Serving humans will require a capability to detect many different objects and to learn about them. Ideally, robots should be able to learn about objects without constant or dedicated supervision, but rather like children do, during interaction with adults and by manipulating objects [1].

Objects appearances can be learned through observation. However, more complete objects representations are required, when a robot needs to exploit objects for accomplishing tasks. This information can be essentially retrieved through active object exploration [2]. Manipulation provides an opportunity to gather an object appearance from different viewing angles and scales by turning the object and approaching to a camera. However, during manipulations, objects are often partly covered by a robot's or human hand, and thus the ability to distinguish between features that belong to the robot, the human, and the manipulated object, is crucial. This paper focuses on this issue: we propose an approach to enhance learning through object manipulation and categorization of visible entities into robot's parts, human parts, and manipulable objects. The interplay of the implemented modules is shown in Fig. 1.

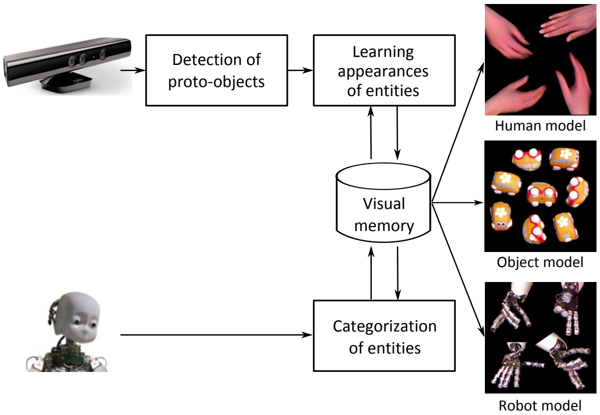
Detection of proto-objects
Learning appearances of entities
Visual memory
Categorization of entities
Human model
Object model
Robot model

Fig. 1. The main modules of the proposed system.

Self-identification has been used in various applications, as it endows the robot with a better control of its body [3], it facilitates interaction with humans and objects. The ability to distinguish between several individuals or sources of motion also gives the robot an opportunity to understand the dynamic of its environment and to interact with several persons [4].

Among the variety of robot self-recognition methods, most algorithms are based on local approaches or prior knowledge. Some strategies impose restrictions on the change of motors configurations during self-recognition. Others exploit a predefined appearance of the robot's body or a predefined pattern of the robot's motion that simplifies the self-identification [3]. Since these techniques are not independent of the appearance of the robot's body and behavior, they cannot be easily generalized over new end-effectors and they cannot recognize robot's parts extended by grasped tools, that would be useful to increase the robot capabilities.

Following a developmental robotics approach, we take inspiration from the sensorimotor developmental stage in humans. Observations show, that at the beginning of life, infants learn about own body through simple repetitive movements, and then spend a lot of time by exploring surrounding objects through interaction [5]. These exploratory actions become effective, once toddlers learn to control and recognize their own body [6]. Our preliminary experiments investigating this issue with the iCub robot are presented in [7].

In this paper we propose a self-identification, categorization, and learning method which is able to differentiate and to memorize appearances of objects, humans, and robot's parts. The algorithm builds upon our previous learning approach [8] and introduces new elements integrating the robot's

[1] U2IS, ENSTA ParisTech - INRIA FLOWERS Team, 828, Boulevard des Marechaux, 91762 Palaiseau Cedex, France firstname.lastname at ensta-paristech.fr

[2] ISIR, UPMC, Paris, France serena.ivaldi at isir.upmc.fr

actions into the learning process and improving the final learning performance. Our algorithm does not require prior knowledge about the robot or objects appearances, robot's body model (kinematics or dynamics), nor the functional description of its joints, and is thus easily adaptable to different robots.

The paper is organized as follows: Section 2 gives a brief overview of the related work on robot self-discovery and its applications; the proposed approach is detailed in Section 3; the performed experiments and their evaluation are reported in Section 4; the last Section is devoted to conclusions and future work.

II. RELATED WORK

Self-identification has been performed using several approaches. It can be achieved based on a known robot's appearance, or a predefined pattern of the robot's motion [3]. The identification of a robot's hand can be also based on temporal contingency, for example, by learning the time delay between the initiation of the robot's movements and the emergence of its parts in the visual field, as proposed in [4]. However, methods based on time delay are often limited to one active motion source at a time.

The identification of robot's parts without prior knowledge can be based on correlation between the proprioceptive and sensory information. This information can be analyzed during head-arm movements, as performed in [9]. The authors analyse the speed of visual motion and of the robot's joints to recognize the robots arms and learn its appearance.

A system discovering robot's hands during natural interaction with a human is presented in [10]. Mutual information is used to identify which salient region of the visual space can be influenced by the robot's actions: the algorithm analyzes the visual input and proprioceptive sensing. Since it is designed to detect humans and robot's parts, it focuses on regions that are close to the sensor and move fast.

Assuming knowledge of the robot's body, several studies exploit the robot's actions for object exploration. The decomposition of scene into objects by means of interactive actions is proposed in [11]. In [12], perception and interaction are integrated for autonomous acquisition of kinematic structures of rigid articulated objects. The interactive learning of objects features and object-specific grasping knowledge is performed in [13]. Robots actions are also used to improve object recognition in ambiguous situations. Having several similar objects, interaction can be used to turn one object into a representative perspective that allows to recognize it [14].

In our approach, we do not focus on the selection of a particular action to act on objects, or use of actions for object segmentation; we rather attempt to learn objects appearances in between actions and during manipulations, while the objects are grasped. As a consequence, the discrimination between manipulated objects, the robot's and human parts is fundamental.

III. PROPOSED METHOD

Our approach detects proto-objects as salient regions of the visual space, incrementally encodes their appearance, and associates them with physical entities. The learning algorithm is based on our previous work on object learning through observation [8], but it has been improved with Bayesian filtering in order to enhance temporal coherency of object recognition and enhanced with a capability of categorization and interactive learning. Entities are classified into robot's hands, human hands, and manipulable objects. The pose of each object entity, its dimensions, and its localization in the robot's space are estimated in order to plan the robot's actions. Finally, the object learning is improved through manipulation using the outcome of categorization.

As input data, we use color and depth images from a RGB-D sensor (Kinect) and robot's motors states. The complete experimental setup will be described in section IV.

A. Segmentation of the visual space

The visual attention in our approach is based on motion; we therefore begin proto-object detection by estimating moving regions by image differencing. Among all moving regions, we ignore whose located far from the robot according to the constraints of the reachable area. In remaining regions, GFT-points are extracted and grouped into clusters of coherently moving points. Each cluster is considered a proto-object and tracked in time. The contours of proto-objects are refined based on the depth variation of the visual field. The processing steps are detailed in [8] and summarized in Fig. 2.

B. Robot actions

Before interaction, we localize proto-objects in the operational space of the robot, estimate their orientations and dimensions. By retrieving the depth information from the RGB-D sensor and processing it as a point cloud, we compute each proto-object's 3D position relative to the sensor before transforming it to the operational space. The proto-object's axes orientations are obtained from eigenvectors and eigenvalues of the covariance matrix of the proto-object's points giving three orthogonal reference directions for which we compute the proto-object's dimensions.

Since this study is aimed at learning objects appearances, the robot should perform actions that help to explore different object perspectives. Thus, we use both simple actions, like $reach$, $push$, $take$, and more complex manipulations, $TakeLiftFall$ and $TakeObserve$, that are aimed at revealing new object perspectives. Both manipulations are composed from a sequence of action primitives. $TakeLiftFall$ includes reaching an object, taking it, lifting, and releasing that turns the object into a random perspective, when it falls on the table. $TakeObserve$ consists of reaching an object, taking it, turning, approaching to the camera, and returning to the table; during this manipulation, the robot perceives several object perspectives and its visual details.

C. Object model learning

The proto-objects appearances are learned incrementally based on the algorithm presented in [8]. Our system acquires all information iteratively by analyzing low-level image

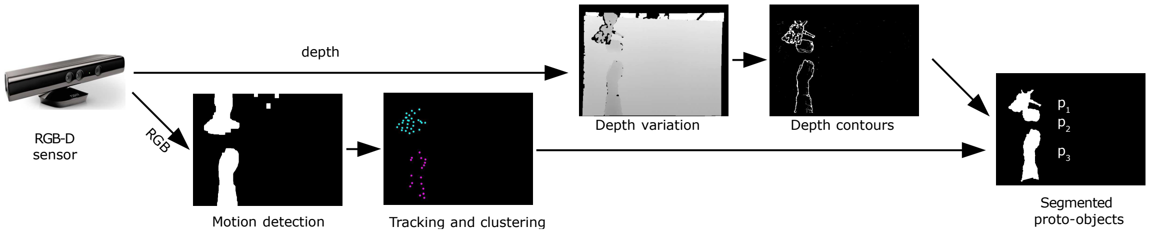

Fig. 2. The segmentation of the visual space into proto-objects p_0, p_1, p_2. See [8] for a complete description.

features and synthesizing them into higher-level representations. As low-level features, we extract SURF points [15] and colors of superpixels [16] that correspond to nearly homogeneous image regions segmented with some regularity. Mid-features are constructed as pairs and triples of low-level features nearest in the visual space. All extracted features are quantized into vocabularies of visual words. Since an entity appearance can vary between its perspectives, we learn its model as a set of views $E_i = \{v_j\}$, each view being encoded by the occurrence frequencies of its mid-features $v_j = \{m_k\}$.

In [8], views are recognized through a voting method based on TF-IDF (Term-Frequency - Inverse-Document Frequency [17]) of mid-features and a maximum likelihood approach:

$$L(v_j) = \sum_{m_k \in v_j} tf(m_k) idf(m_k), \tag{1}$$

where $tf(m_k)$ is the frequency of the mid-feature m_k, and $idf(m_k)$ is the inverse view frequency for this mid-feature.

Since several objects can have similar views, we introduce a Bayesian filter that improves temporal consistency of recognition between consecutive images therefore reducing potential confusion between objects. The probability of recognizing a view is estimated recursively based on its likelihood, its probability computed in the previous image, and its tracking:

$$p_t(v_j) = \eta L(v_j) \sum_l p(v_j|v_l) p_{t-1}(v_l), \tag{2}$$

where η is the normalization term; $L(v_j)$ is the current likelihood of the view v_j; $p_{t-1}(v_l)$ is the probability of the view v_l computed in the previous image; $p(v_j|v_l)$ is the probability that the view v_j appears, when the view v_l was recognized in the previous images. This probability is fixed to 0.8 when $v_j = v_l$, and otherwise $0.2/N_v$ with N_v being the total number of views.

The recognized view is then associated with a physical entity. If the entity tracking from previous image was successful, the view is associated to the same entity. When tracking fails, the current entity is recognized through a maximum likelihood approach similar to the view recognition but based on the occurrence frequency of views among entities:

$$L(E_i) = tf(v_j) idf(v_j), \tag{3}$$

where $tf(v_j)$ is the frequency of the view v_j, and $idf(v_j)$ is the inverse entity frequency for the view v_j.

Since our experiments are based on object manipulation, it is important to recognize connected physical entities moving together, while the robot or the human interacts with an object. For this purpose, we use a double-check recognition. In the first stage, the most probable view is identified. In the second stage, features that don't belong to the most probable view (see Fig. 3) participate in the voting method again to identify a second possible view. Thus, each moving region of the visual space is recognized either as a single entity or two connected entities. Since objects are partly covered by hands during manipulations, the double-check recognition allows to prevent erroneous updates of objects models with hand features. The information about connected physical entities is also used by the categorization module described in the Section III.D and during interactive object learning presented in the Section III.E.

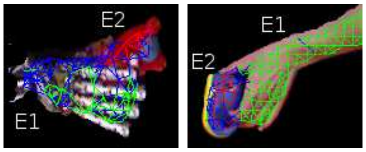

Fig. 3. Recognition of connected views: the mid-features (in this case, pairs of superpixels) found in the most probable view are shown by the green color, the mid-features found in the connected view are red, and the rest of extracted mid-features are blue.

D. Categorization

The categorization procedure is aimed at identifying the nature of physical entities detected in the visual space, while the robot learns objects through interaction with a human partner. First, the parts of the robot's body are discriminated among all entities, and then, the rest of single entities are distinguished either as a human part or a manipulable object category. As a result, each entity is associated with one of the following categories (see Fig. 4): a robot c_r, a human c_h, an object c_o, an object grasped by the robot c_{o+r}, an object grasped by the human c_{o+h}, or unknown c_u category that will be identified later, when more statistics is gathered.

1) Robot self-identification: The robot's body identification is based on mutual information (MI) between visual data and proprioception. As proprioceptive information, we analyze the robot's arm and torso motors states. We acquire states of the following arm joints (see Fig. 4): shoulder

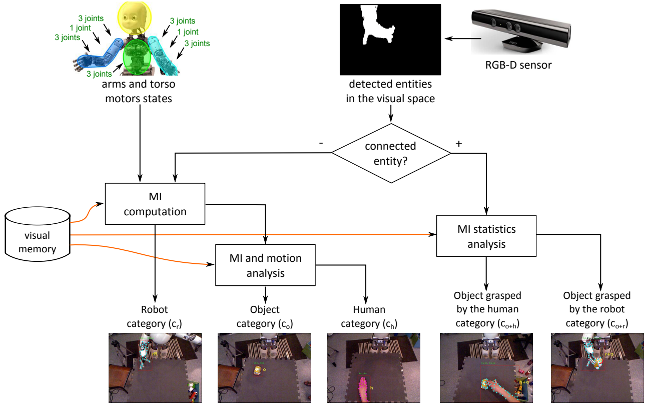

Fig. 4. The categorization algorithm: single entities are categories as c_r, c_h or c_o based on mutual information between the visual data and proprioception and statistics on entities motion; connected entities are categorized as c_{o+r} or c_{o+h} based on the entities categorization statistics throughout the whole experiment

(pitch, roll, and yaw), elbow, wrist (pronosupination, pitch, and yaw) and torso joints (pitch, roll, and yaw). Finger joints are not considered, since their movements do not produce a significant visual displacement of the hand.

The visual space is quantized regularly by applying a grid (12x10) producing 120 visual clusters. The position of each physical entity is quantized to the closest visual cluster. Each time a new image is acquired from the visual sensor, we acquire the robot's arms and torso joints values. The joints values are incrementally quantized into a vocabulary of arm-torso configurations, where each entry is encoded as a vector of joints values. During quantization, if the minimal L2 distance between the current vector of joints values and each vocabulary entry exceeds a threshold, a new configuration is stored in the vocabulary; otherwise, the current vector of joints values is recognized as the closest arm-torso configuration from the vocabulary. In our experiments, we obtain in average 37 arm-torso configurations.

As in [10], MI is used to evaluate the occurrence dependency between the robot's arm-torso configuration A_c and the physical entity localization L_{E_i}:

$$MI(L_{E_i}; A_c) = H(L_{E_i}) - Hc(L_{E_i}|A_c), \qquad (4)$$

where $H(L_{E_i})$ is the marginal entropy, and $Hc(L_{E_i}|A_c)$ is the conditional entropy computed in the following way:

$$H(L_{E_i}) = -\sum_{l} p(l_{E_i}) log(p(l_{E_i})), \qquad (5)$$

$$Hc(L_{E_i}|A_c) = -\sum_{a_c} p(a_c) \sum_{l_{E_i}} p(l_{E_i}|a_c) log(p(l_{E_i}|a_c)), \qquad (6)$$

where $p(l_{E_i})$ is the probability of the entity localization l_{E_i}, $p(a_c)$ is the probability of the arm-torso configuration a_c, and $p(l_{E_i}|a_c)$ is the probability of the entity localization l_{E_i} given the arm-torso configuration a_c.

Since we change the appearance of the robot's hands during experiments, $MI(L_{E_i}; A_c)$ is estimated for each robot's arm and for each physical entity. Thereby, the robot category c_r can be associated with several entities that correspond to different appearances of the hand (for example, with and without wearing gloves); while views of each entity describe the hand appearance in different postures (see Fig. 5).

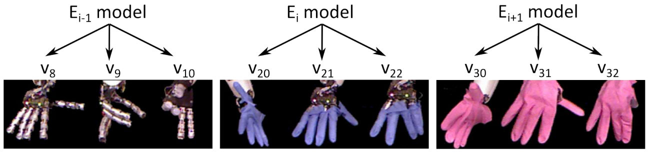

Fig. 5. The representation models of three entities that correspond to different appearances of the robot's hands.

The threshold identifying the robot category is selected empirically by analyzing the MI distribution for robot's and non-robot's parts on a small labelled database. If MI is higher than $th_r = 40\%$, the physical entity is identified as a robot category c_r; otherwise, its category is identified according to the algorithm of the following section.

2) Discrimination of human and object categories: The discrimination between human parts and manipulable objects is based on statistics on entities motion: human parts often move by themselves while objects are static most of time, and they are rather displaced by the robot or the human.

Since our vision module is able to detect and to categorize connected entities moving together, we identify objects during manipulations based on the statistics of their simultaneous motion with entities categorized as robot's parts. During the experiment, we count the number of times each entity E_i moves alone as a non-robot category, and the number of times the same entity moves connected to a robot's entity and estimate the associated occurrence frequencies:

- $f_s = \frac{N_{c_{E_i} \neq c_r}}{N_{c_{E_i}}}$ is the occurrence frequency of a non-robot's entity moving alone,
- $f_c = \frac{N_{c_{E_i}, c_{E_{i2}} = c_r}}{N_{c_{E_i}, c_{E_{i2}}}}$ is the occurrence frequency of an entity moving together with a connected entity E_{i2} categorized as a robot's part.

Since objects usually do not move alone, the frequency f_s should be low and f_c should be high for the object category. Therefore, a non-robot's entity is identified as:

- the object category c_o, if $f_c > th_o.c.$ and $f_s < th_o.s.$;
- the human category c_h, otherwise.

Gathering these statistics require the identification of the robot hand category c_r, therefore all entities are temporarily associated with the unknown category c_u before c_r is identified. Once the robot's body is identified, all single entities are categorized as c_o, c_h, or c_r. In the case of connected entities, the category of each individual entity is retrieved from the categorization statistics and the connected entity is categorized as an object grasped by the robot category c_{r+o} or an object grasped by the human category c_{h+o}.

E. Object model update during interaction

The outcome of the categorization module is used to improve object learning during manipulation. The interaction with an object starts when the robot detects an object entity in a reachable distance. In case of a successful grasp, the model of the grasped entity E_g is updated during manipulation. This is a kind of self-supervision, where the object is supposed to be the same during manipulation.

The perceptual system continuously detects entities in the visual space and categorizes them. In the case of detecting connected entities with one entity identified as a robot category, the categories of both connected views are verified. We link each connected view with a set of physical entities $\{E_i\}$ that have this view in their models. The category c_{E_i} of each entity is retrieved from the categorization statistics, and each connected view is identified as:

- a robot's view, if at least one linked entity is identified as the robot category ($\exists i, c_{E_i} = c_r$);
- a non-robot's view, if none of linked entities is identified as the robot category ($\forall i, c_{E_i} \neq c_r$).

If during manipulation, a proto-object is identified as a robot's view connected to a non-robot's view,the manipulated entity model is updated with the non-robot's view. If a proto-object identified as a robot's view contains a large amount of features that do not correspond to this entity, a new view is stored with these features. If this newly created view is identified again later, it will be added to the manipulated entity model. Therefore, interactive learning allows to update the object model with both newly created and recognized non-robot's views.

IV. EXPERIMENTS

The proposed approach is evaluated on an iCub robot interacting with a human partner, as demonstrated in Fig. 6a, and manipulating objects, as shown in Fig. 6b. Objects used in the experiments are shown in Fig. 7.

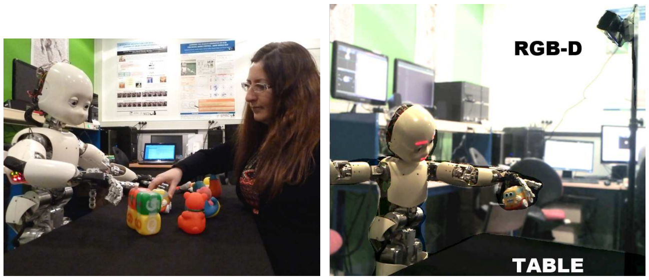

Fig. 6. The context of the experiments: a) learning through observation; b) learning through manipulation.

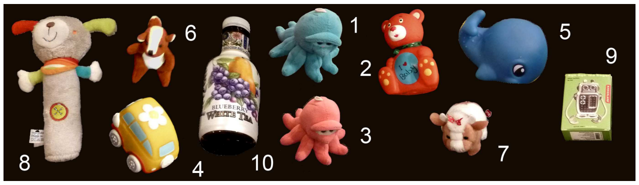

Fig. 7. Objects used in the experiments.

We design experiments for two purposes: first, to evaluate the categorization algorithm and then to analyse the accuracy of objects learning through manipulation and to compare it with the results of learning through observation.

A. Camera calibration

In our experiments, the visual input is acquired from an RGB-D sensor mounted above the robot (see Fig. 6b). This sensor is chosen due the precision of depth data compared to stereo vision. Since in our scenario, the robot performs actions in its operational space, the visual sensor is calibrated with respect to the robot, like described in [7]. In this procedure, a calibration pattern is placed on the table and the robot moves its hand to the origin of the pattern in order to acquire its position in the operational space $H_{pat \to rob}$. The OpenCV library is used to estimate the sensor position relative to the pattern $H_{sen \to pat}$, and the transformation matrix from the target to the robot's space is computed:

$$H_{sen \to rob} = H_{pat \to rob} \times H_{sen \to pat}.$$

B. Evaluation of categorization

In this experiment, a human manipulates objects and produces simple hand movements in the visual field of the robot. The robot performs simple actions, like reach, take, push, and manipulations with and without objects, as described in the Section III.B. The self-identification method is evaluated based on the robot's hands positions estimated by the forward kinematics model.

During evaluation, the categorization module was able to identify the robot's hand within first 10 seconds of its motion in the visual field. The average self-recognition rate was about 98.2%. Our self-identification method is also evaluated with changing the robot's hand appearance by wearing colored gloves (see Fig. 5). The system has shown to be independent on the robot's hand appearance and to recognize 98.1% of the robot's hands in the blue gloves and 98.0% of the robot's hands in the pink gloves. The slightly lower self-recognition accuracy in the case of changing the hand appearance can be explained by a large sizes of the gloves that reduce visibility of hand motion.

The system's ability to identify an object category is evaluated in an interactive scenario, while the robot is asked to interact with entities detected at a reachable distance. As shown in Fig. 8, each object has been successfully identified during within 5-10 seconds of interaction with it. Human parts have been correctly identified in 89% of images.

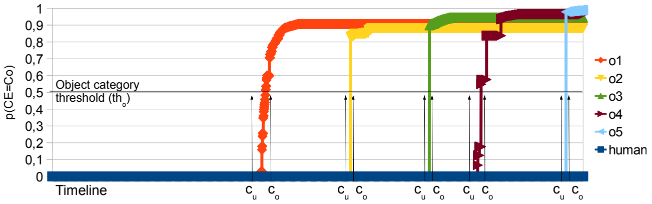

Fig. 8. Identification of five objects based on their probability $p(c_E = c_o)$ of being an object category c_o; each object is marked as an unknown category c_u, when it appears, and as c_o, when it is identified.

C. Evaluation of object learning

We evaluate the accuracy of objects learning through interaction and compare it with the results of learning through observation. During observation, a human demonstrates objects to the robot (about 700 images per object). Then, during manipulation, the robot performs $TakeLiftFall$ action (about 800 images per object) in order to improve its knowledge about objects appearances.

Since our experiments are based on interaction with objects, it is difficult to evaluate the system using existing image databases. Thus, we created a database of 50 images for each object shown from different perspectives. This database is processed after each experiment in order to estimate the object recognition rate based on the number of times an object is identified as its most frequently associated entity.

Learning through manipulation improved the recognition rate for several objects compared to the results of learning through observation (see Fig. 9). This improvement slightly depends on the robot's hand appearance; the best results have been achieved with the robot's hand appearance the most different from all objects appearances, i.e. without wearing gloves. Gloves produce a larger occlusion of object features, making it less visible and leading to less updates of object models and smaller learning improvement.

Using only observation, several objects whose appearance significantly varies between perspectives are associated with multiple physical entities. It occurs when the human partner takes an object out of the visual field while demonstrating different perspectives, making it impossible to track the object and therefore to associate all its views with a single entity. For these objects (O_2, O_4, O_5, O_8, O_9), learning during manipulation has been especially useful as several entities created during observation have been merged into a single entity during interactive learning, thus leading to better object recognition. Moreover, the system was able to memorize new views while manipulating objects O_1, O_6, O_8, thus improving the informativeness of objects models.

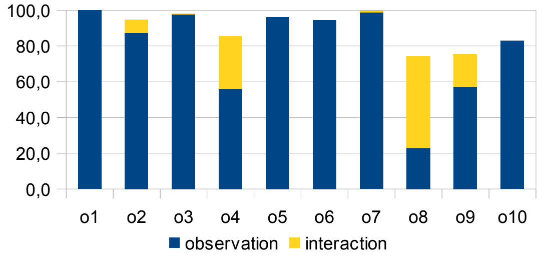

Fig. 9. Object recognition rate (with initial robot's hands appearances): the results after objects observation are shown by the blue color, and the improvement after manipulation is shown by the yellow color.

V. CONCLUSION AND FUTURE WORK

The proposed developmental approach allows a robot to explore its close environment in purely unsupervised way, to identify its body and to categorize other visible physical entities as human parts or manipulable objects. Based on these categories, it is possible to learn objects through observation and to improve their visual models through manipulation.

Important aspects of our model are its capacity to extract new information about an object during and in between manipulations and its adaptability to the modification of the robot's appearance. The system works online and gathers all information in an incremental manner.

Future work will include the use of weak supervision by integrating the audio information in our system. We plan to take advantage of naming objects, like in infant directed speech, in order to learn objects names and to improve object recognition in more complex interactive scenarios.

ACKNOWLEDGMENT

Our work is performed as a part of MACSi[1] (Modele pour l'apprentissage du comportement sensorimoteur d'iCub) project supported by the French ANR program (ANR 2010 BLAN 0216 01).

REFERENCES

[1] S. Perone, K. Madole, and L. Oakes, "Learning how actions function: The role of outcomes in infants' representation of events," *Infant Behav Dev*, vol. 34(2), p. 351362, 2011.

[2] L. M. Oakes and H. A. Baumgartner, "Manual object exploration and learning about object features in human infants," in *IEEE Int. Conf. on Development and Learning and Epigenetic Robotics (ICDL)*. IEEE, 2012, pp. 1–6.

[3] L. Natale, "Linking action to perception in a humanoid robot: A developmental approach to grasping," *PhD diss., Univ. Of Genoa*, 2004.

[4] P. Michel, K. Gold, and B. Scassellati, "Motion-based robotic self-recognition," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, vol. 3. IEEE, 2004, pp. 2763–2768.

[5] J. Piaget, *Play, dreams and imitation in childhood*. London: Routledge, 1999.

[6] 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, 2009.

[7] 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 *IEEE Int. Conf. on Humanoids*. IEEE, 2012.

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

[9] R. Saegusa, G. Metta, G. Sandini, and L. Natale, "Action learning based on developmental body perception," in *IEEE Int. Conf. on Industrial Technology (ICIT)*, 2013.

[10] 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.

[11] H. van Hoof, O. Kroemer, H. B. Amor, and J. Peters, "Maximally informative interaction learning for scene exploration," in *IEEE Int. Conf. on Intelligent Robots and Systems (IROS)*, 2012.

[12] D. Katz, A. Orthey, and O. Brock, "Interactive perception of articulated objects," in *12th Intern. Symp. of Experimental Robotics*, 2010, p. 1.

[13] D. Kraft, R. Detry, N. Pugeault, E. Baseski, F. Guerin, J. H. Piater, and N. Kruger, "Development of object and grasping knowledge by robot exploration," *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 4, pp. 368–383, 2010.

[14] B. Browatzki, V. Tikhanoff, G. Metta, H. Bulthoff, and C. Wallraven, "Active object recognition on a humanoid robot," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*. IEEE, 2012, pp. 2021–2028.

[15] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Comput. Vis. Image Underst.*, vol. 110, no. 3, pp. 346–359, Jun. 2008.

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

[17] 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.

[1] www.macsi.isir.upmc.fr