# Discovering Affordances Through Perception and Manipulation

Omar Ricardo Chavez-Garcia, Pierre Luce-Vayrac, Raja Chatila

# Discovering Affordances Through Perception and Manipulation

R. Omar Chavez-Garcia, Pierre Luce-Vayrac and Raja Chatila

***Abstract*— Considering perception as an observation process only is the very reason for which robotic perception methods are to date unable to provide a general capacity of scene understanding. Related work in neuroscience has shown that there is a strong relationship between perception and action. We believe that considering perception in relation to action requires to interpret the scene in terms of the agent's own potential capabilities. In this paper, we propose a Bayesian approach for learning sensorimotor representations through the interaction between action and observation capabilities. We represent the notion of affordance as a probabilistic relation between three elements: *objects*, *actions* and *effects*. Experiments for affordances discovery were performed on a real robotic platform in an unsupervised way assuming a limited set of innate capabilities. Results show dependency relations that connect the three elements in a common frame: affordances. The increasing number of interactions and observations results in a Bayesian network that captures the relationships between them. The learned representation can be used for prediction tasks.**

## I. INTRODUCTION

Scene understanding has traditionally been addressed as a process of observation. Even if active vision was introduced by Krotov & Bajcsy in [1], this was to gain more information through exploitation of different viewpoints. Considering perception in relation to action requires to interpret the scene in terms of the agent's own potential activity. It is increasingly acknowledged by psychologists, neurocientists and roboticists that perception and action are parts of an interactive, developmental and integrated process [2], [3].

Reasoning jointly on perception and action requires self-localization with respect to the environment. Hence, developing sensorimotor representations and not just environment representations puts the robot in the center of the perceptual process, and provides for a link between self-awareness and situation-awareness.

The representative features from perception and action are typically defined in different feature spaces, making it difficult to find relationships over different datasets, even when they are semantically related. The need for a fusion approach that encodes sensorimotor correlations, without losing frame-related information, becomes key for a developmental approach.

In this paper, we propose a Bayesian approach for learning sensorimotor representations in terms of affordances represented as probabilistic relations between three elements of the perception-action process: *objects*, *actions* and *effects*. Our learning approach relies on acquired information from interaction, and the innate robot knowledge for the elements of the perception-action process. These innate capabilities are kept minimal and will be extended with experience.

Figure1 shows a pipeline of the proposed affordance learning approach. First, we develop a sensory perception process to extract, from RGB-D data, salient elements which are hypothesized as *objects*. Then, from a set of innate basic *actions*, we select an action to execute on an object. Innate *effect* detectors are triggered in case a change in the perceptual or proprioceptual components is present during the interaction. Finally, sensorimotor representations emerge from affordances through the stochastic fusion of perception and action components.

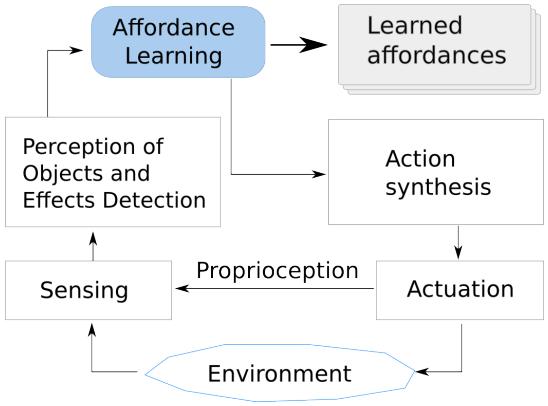


Fig. 1. Pipeline of the proposed affordance learning approach.

The rest of the paper is organized as follows. In section II we discuss related work. Section III describes our perception approach. In section IV, we define our sensorimotor representation, its elements and the fusion approach for learning affordances. Section V details our experimental setup and the results evaluation. Conclusions and perspectives are presented in section VI.

## II. RELATED WORK

Robotic autonomy aims at performing several tasks inside a dynamic environment. Robots need to be able to identify and learn novel objects, discriminate and recognize learned objects, perceive their own dynamics and relate which actions can be performed with certain objects in order to comply with required tasks.

In recent years, several works have followed developmental ideas for exploring and learning robots' environment [4], [5], [3]. This is generally carried on via a cycle of

R. Omar Chavez-Garcia, Pierre Luce-Vayrac and Raja Chatila are with the Institut des Systèmes Intelligents et de Robotique (ISIR), Sorbonne Universités, UPMC Univ Paris 06, CNRS, 75005 Paris, France. {chavez,luce-vayrac,raja.chatila}@isir.upmc.fr

exploration-manipulation which is initialized with a collection of minimal knowledge and innate capabilities.

Deterministic approaches such as [6] and [7], propose learning affordances for pre-defined categories. Here the relations between perception and action were not discovered exclusively by exploration and relied on supervised classification between predefined categories. In [8] and [9] the categorization between effects is found using unsupervised clustering in the effect space. However, the categorization was based only on the mapping between objects and effects, leaving the action out of the process. Moreover, object representation of the aforementioned work was related to object recognition rather than object description.

Imitation-inspired and probabilistic-powered methods have been explored in [10] and [11]. The latter work extracts *knowledge* through exploration from other *knowledge*, and uses Bayesian Belief Networks to exclusively associate motor commands and forward models. The former approach, uses relational dependency networks to learn joint probability estimates regarding the effect of sensorimotor features on the predicted quality of desired behaviors. This approach allows robots to determine which features in the world are relevant to certain motor commands. In both works, the notion of object is hardwired focusing only on the actions-effects association.

An emerging field has categorized these works as *Interactive Perception* because they exploit sensory signals due to interaction which would otherwise not be present. In addition, these systems, in a minor or major degree, leverage knowledge about regularities found in the combined frame $Sensor \times Action \times time$ [12]. Works like [13] and [14] propose stochastic methods to discover object properties by interaction. Although their results show important improvements in object segmentation and action planning, they do not consider the action component as a relevant element of the sensorimotor learned model.

In our work, we follow a bottom-up approach that starts from low-level data from sensors and actuators, up-to learning relations between higher-level representations. The probabilistic nature of the work maintains the uncertainty characteristic of the perception-action cycle. Our sensorimotor representation encodes, through the learning of affordances, effects, objects and actions in the same feature space. It enhances works such as [5] and [3] by including an information-based methodology to find the relations in the combined feature space instead of relying on a preferred-relation list. Following the categorization proposed by Bohg et al. [12], our work can be considered a multimodal Interactive Perception approach with given priors on the robot dynamics, and on the observations. It has as goals automatic object segmentation, estimation of intrinsic object parameters, sensorimotor learning, and eventually semantic categorization.

## III. SENSORY PERCEPTION

Usually, segmentation algorithms only consider low-level information from the image or point cloud. Recent semantic segmentation methods take advantage of high-level object knowledge to help disambiguate object borders [15], [16].

### A. Over-segmentation approach

Supervoxels are formed by over-segmenting a 3D image into small regions based on local low-level features, reducing the number of nodes which must be considered for segemen-tation.We implemented a 3-D version of the Voxel Cloud Connectivity Segmentation presented in [17], which takes advantage of 3D geometry provided by RGB-D cameras to generate supervoxels evenly distributed in the observed space. This uses a seeding methodology based on 3D space and a flow-constrained local iterative clustering which uses color and geometric features. Due to ensuring strict partial connectivity between voxels, this algorithm guarantees that supervoxels cannot flow across boundaries which are disjoint in 3D space.

Supervoxels features are represented by 39-dimension vector composed of spatial coordinates $(x, y, z)$, color information (Lab color space), and 33 elements from an extension of the Point Feature Histogram: $F = [x, y, z, L, a, b, FPFH_{1..33}]$ [17]. This offers a multi-dimensional pose-invariant representation based on the combination of neighboring points. For each supervoxel, in an outward direction, we calculate a normalized spatial distance $D_s$, a normalized color distance $D_c$ and the distance in the FPFH space $D_{HIK}$ [17]. If the distance is the smallest seen, this voxel and its neighbors (in the adjacency graph) become part of the supervoxel. The result, as shown in Figure 2, is an over-segmented cloud where each supervoxel (segment) cannot cross over object boundaries that are not actually touching in 3D space. Supervoxels tend to be continuous in 3D space, since labels flow outward, at the same rate, from the center of each supervoxel [17].

### B. Intrinsic clustering

Figure 2 shows how supervoxels are still considered representations of individual patches. A clustering process is needed to group the supervoxels that possibly correspond to the same object without relying on *a priori* information about the number of objects. Regarding the feature representation detailed in Section III-A, we proposed to use the non-parametric technique described in [18] to find the shape of the object hypotheses based on the set of supervoxels.

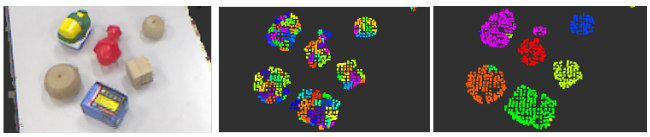

Fig. 2. Results from the sensory perception process. RGB-D cloud of points from the real scenario (left); over-segmentation results from point cloud (middle); results from intrinsic clustering (right).

Figure 2 shows the result of the clustering method. The result of this intrinsic clustering is a set of labels $L_{hyp}(t)$ for a group of supervoxels that represent hypotheses of objects in the current scenario.

### C. Object hypotheses confirmation

The set of generated hypotheses from Section III-B are built only using the sensory data. This means that segmentation issues can appear in the form of incomplete, divided and false segments of real objects in the scenario. We perform a tracking-by-detection approach to reduce the number of false positive segmentations. Only the active segments hypotheses with tracks lengths over a threshold $\tau$ are considered as confirmed objects for our sensory-perception task. Each object is represented by its centroid, which offers a point of interaction ($poi$) for the interaction task.

### D. Features extraction

In this work, we assume that the robot has minimal innate perceptual capabilities that allow it to discretize the environment. These capabilities are related to segmentation. It can differentiate from color values. It has geometrical notion of position, continuity of segments and normal extraction for surfaces. Therefore, the robot can extract higher level features for the description of confirmed objects. By analyzing the segment representing an object, we focus on three main features: $color$, $size$ and $shape$. Transforming from RGB to HSV color model, we extract the dominant hue of the object. Size of the object is obtained from the distance between the start and end of the largest segment of the cluster representing the object. Four-dimensional templates are used to select the form of the object from a set of fixed three-dimensional forms: *cube*, *cuboid*, *sphere*, *irregular*. Our object description allows for expanding the set of perceptual features, e.g., histogram of shapes.

## IV. SENSORIMOTOR LEARNING

In addition to the perceptual information, object manipulation allows the robot to learn sensorimotor correlations between the sensor inputs fused in the object descriptions $O$, robot basic actions $A$ and the salient changes represented by the effects $E$. Starting from built-in actions, the evolution of the environment is captured by perception through the information provided by effect detectors, e.g. object movement detection and proprioceptive feedback. The goal is to learn from regularities in the occurrences of elements in $O$ and $E$ when an action $a_i \in A$ is triggered.

### A. Actions

Making an analogy to a newborn's rough motor abilities [19], we assume that the robot is built with a set of basic motor capabilities, which we call actions. In this set of basic actions $A = \{a_1, ..., a_n\}$ , each action can be defined as follows:

$$a_i(V^*, \gamma, \sigma_{a_i}), \tag{1}$$

where $V^*$ represents the desired value for the robot controlled variables $V$, $\gamma$ is its proprioceptive feedback and $\sigma_{a_i}$ represents the parameters for the particular action $a_i$.

Hypotheses obtained from sensory perception provide points of interest in the perceptual frame identifying objects. These points are used as targets for the actions approaching to the object through perceptual servoing.

The interaction focus of our work is on object manipulation, therefore our proposed set of actions $A$ is composed of 4 actions. Lean-toward ($lean\_t$) moves toward the current position of the end effector in a constant velocity fashion. Poke ($poke$) behaves in a similar way as $lean\_t$ but using a constant acceleration movement. Open ($open\_g$) and close gripper ($close\_g$) are self-explanatory. Due to the general affordance learning goal of our approach, we decided to fix the parameters $\sigma_{a_i}$, i.e., velocity for $lean\_t$, acceleration for $poke$, and force for $close\_g$. Information provided by $\gamma$ includes the joint and force values of the actuators and the state of the end effector of the robot.

### B. Effect detectors

We consider an effect as a correlation between an action and a change in the state of the environment, which includes the agent itself. For example, when a robot interacts with an object, it can perceive changes related to the position of the object, proprioceptive values of the actuators and feedback from the end effectors. An effect is an important element in our sensorimotor fusion and its detection (or lack thereof) plays the role of common ground for perception and action frames. The robot's innately detectable effects are divided in two groups: perceptual-based (object's linear movement); and proprioceptive-based (end effector linear force, distance between gripper's fingers and effector's linear movement).

### C. Affordance learning

An affordance is an *acquired* relation between an effect $e$ in $E$, a capability (in our case an action) in $A$ over an entity in $O$. One can state that when an agent $g$ applies its capability $a$ over an entity $o$, an effect $e$ is generated [20]. From an agent's perspective, from now on the robot, the $i_{th}$ affordance is defined as follows:

$$\alpha_i = (e_l, (o_k, a_j)), \text{ for } e_l \in E, o_k \in O, \text{ and } a_j \in A. \tag{2}$$

Figure 3 shows an example of a relation between an entity $toy$ perceived by the agent $robot$, and the application of its capability $grasp$, implying that there is a potential of generating an effect $grasped$. We can label this relation using its semantic value, $grasp - ability$.

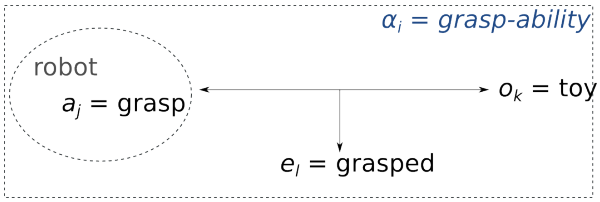


Fig. 3. Representation of an affordance relation labeled $grasp - ability$.

In sections III and IV-A we have defined the three elements mentioned in our affordance definition. We can state our problem as learning the set of affordances for data extracted from $E$, $O$, and $A$.

### D. Structure Learning for affordances discovery

Let us represent the members of the set of elements $E$, $O$ and $A$ as discrete random variables of a Bayesian Network (BN) $\mathcal{G}$. Therefore, we can define these elements as the discretization $E = \{e_l\}$, $O = \{o_k\}$ and $A = \{a_j\}$. Let us assume that through the cycle of perception-interaction we obtained instances of these variables generating a data set $\mathcal{D}$. Our problem of discovering the relations between $E$, $O$ and $A$ can be translated to finding dependencies between the variables in $\mathcal{G}$, i.e., learning the structure of the corresponding BN from data $\mathcal{D}$. Using the BN framework we are capable of displaying relationships between variables. The directed nature of its structure allows us to represent cause-effects relationships. It can handle uncertainty through the established probability theory. In addition to direct dependencies, we can represent indirect causation.

One approach for inducing BN structures from data is the score-based technique, especially for the purpose of probability distribution function estimation. The process assigns a score to each candidate BN that measures how well that BN describes the data set $\mathcal{D}$. For a BN's structure $\mathcal{G}$, its score is defined as the posterior probability given the data $\mathcal{D}$:

$$Sc(\mathcal{G}, \mathcal{D}) = P(\mathcal{G}|\mathcal{D}). \tag{3}$$

A score-based algorithm attempts to maximize this score. Usually, this score is rewritten using Bayes' rule as:

$$Sc(\mathcal{G}, \mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{G})P(\mathcal{G})}{P(\mathcal{D})}, \tag{4}$$

where algorithms only need to maximize the denominator, since $P(\mathcal{D})$ does not depend on $\mathcal{G}$. If we assume a uniform prior over the structures, we can focus only on $P(\mathcal{D}|\mathcal{G})$. Usually, score functions work on the logarithmic space, i.e., $log(P(\mathcal{D}|\mathcal{G}))$. These algorithms select various structures for examination and score them. The structure with the highest score is selected. In this work, we implement an information-based score.

*1) Information compression score:* We can define the score of a BN as the compression rate of the data $\mathcal{D}$ with an optimal code induced by the BN. $\mathcal{D}$ represents the interactions of the robot with values for the variables in $O, A$ and $E$. Using Shannon's noiseless coding theorem, we establish the limits of the compression rate. Therefore, as the number of independent and identical distributed random variables tends to infinity, no compression of the data is possible for a rate less than the Shannon entropy, without losing information.

Bayesian Information Criterion (BIC) is a generalization of the Minimum Description Length (MDL) score, which uses a penalization based on the number of bits needed to compress $\mathcal{D}$, preferring simple BN over more connected and complex ones [21]. We calculate the quality of $\mathcal{G}$ as:

$$\eta(\mathcal{G}|\mathcal{D}) = I(\mathcal{G}|\mathcal{D}) - s(N)|\mathcal{G}| \tag{5}$$

where $I$ is the log-likelihood score that measures the number of bits needed to describe $\mathcal{D}$ given $P(\mathcal{G})$, and $|\mathcal{G}|$ denotes the network complexity. This criterion does not depend on a prior probability distribution on all the possible networks $P(\mathcal{G}|\mathcal{S})$ which is usually a requirement for bayesian scores.

BIC uses a penalization defined as $s(N) = \frac{log(N)}{2}$ to represent the number of bits needed to encode $\mathcal{G}$. In order to increase the likelihood of a structure, we can add parameters, which can result in overfitting. BIC penalizes structures with larger number of parameters.

*2) Search algorithm:* Our implementation, based on the hill-climbing technique for learning BN structures, takes as the inputs values for the variables in $E$, $O$, and $A$ obtained from robot's interaction.

The procedure estimates the parameters of the local probability functions given a BN structure. Typically, this is a maximum-likelihood estimation of the probability entries from the data set, which for multinomial local pdfs consists in counting the number of tuples that fall into each table entry of each multinomial probability table in the BN. The algorithm's main loop consists in attempting every possible single-edge addition, removal, or reversal, making the network that increases the score the most the current candidate, and iterating. The process stops when there is no single-edge change that increases the score. There is no guarantee that this algorithm will settle at a global maximum, but there are techniques to increase its reaching possibilities. We use simulated annealing [22].

## V. EXPERIMENTS AND EVALUATION

The interest of the experiments is mainly in understanding and relating the effects generated by the set of basic actions. Our experiments rely on two assumptions. First, when the robot repeatedly performs a particular action over a particular object, the obtained effect is mostly the same. Second, explicit information regarding the success of an action is not provided; it is obtained through inference over the learned BN. Therefore, our experiments are carried in an unsupervised fashion. Starting with random exploration, we use a motivational system to focus on object hypotheses closest to the end effector to generate (object, action) couples.

### A. Experimental setup

Our Baxter robot (see Figure 4) is equipped with 2 arms with 7 degrees of freedom and torque sensors. One electrical gripper is attached to each arm. Additionally, a Microsoft kinect sensor captures RGB-D data which is used for the sensory perception. For the environment interaction we use the left arm and its gripper.

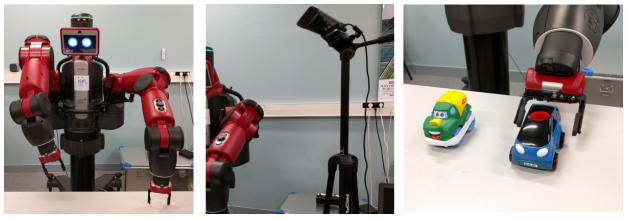

Fig. 4. Experiment setup. Baxter robotics platform (left). Kinect sensor (middle). Subset of objects of interest (right).

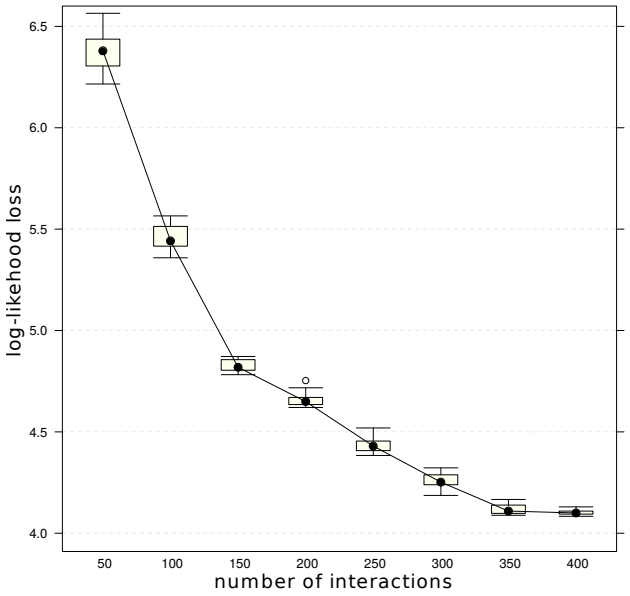


Fig. 5. 10-fold cross validation evaluation of the learned structure. x-axis represents the number of instances (robot interactions) and y-axis accounts for the log-likelihood loss function.

Our training set was generated autonomously by the robot's perception-action interaction. Five objects were used for dataset generation, highlighting the variance in their perceptual information and in their effects with relation to the action set, e.g., objects with different perceptual descriptions and similar expected effects, different expected effects for similar actions. Learning of affordances, as described in section IV-D was done online using data instances obtained periodically. We discretized the values of each effect, object and action variables. A video demonstration of the experimental setup is found at `https://cloud.isir.upmc.fr/owncloud/s/eYof23AtCqHtY4J`.

### B. Experiment results

In figure 5, we present the evolution of the network seen from the point of view of the log-likelihood loss. These results come from the 10 fold cross validation approach over the current set of data. We can see how the expected loss is reduced with the number of interactions. This means that the learned structure is generalizing better for related scenarios. Although most of the network relations evolve with the number of tries, there are some dependency connections that are confirmed in early stages of the experiment (see figure 6).

Figure 6 shows some examples of the relations learned by our approach during the evolution of the experiment. The first relation was learned from the beginning of the experiments. It shows a dependency between the variables representing the perceptual information. The second relation shows a strong casual dependency of the *state of the gripper* ($g\_state$) with respect to the actions *open* and *close gripper* ($open\_g$, $close\_g$). Finally, the third relation shows an example of an affordance relation. It connects the perceptual nodes with the action *lean toward*($lean\_t$) and the effect $poi\_obj\_mov$ which indicates a movement of a point of interest from an

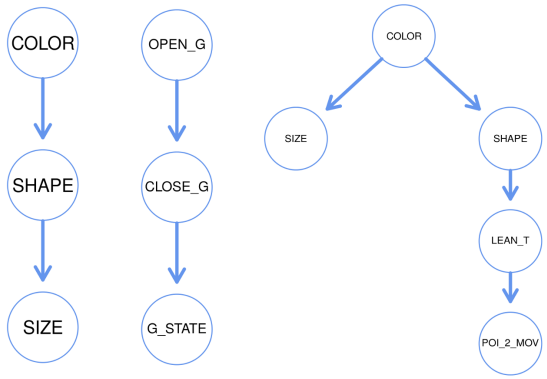


Fig. 6. Examples of relations learned with the proposed approach. Left relation appeared at 30 interactions, middle relation since the iteration 50, right relation appeared at 150 iterations.

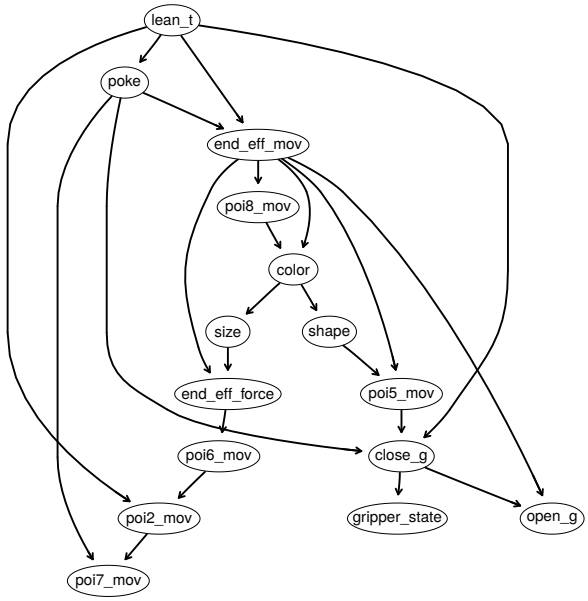


Fig. 7. Learned structure of the BN obtained by our affordances learning method after 400 interactions.

object perceived by the robot. Figure 7 shows the complete learned BN after 400 interactions. Some of the relations shown in figure 6 are kept from early interactions.

Based on the posterior classification error, we perform a 10-fold cross validation scheme to evaluate the performance of the learned network. Table I shows the prediction error for each variable in the learned BN of figure 7. Prediction is computed w.r.t. a set of variables in the BN using likelihood weighting to obtain Bayesian posterior estimates. These results show how robust is the learned network to perform inference with novel data. Prediction error values for variables not present in table I are equal to 0.

Using probabilistic inference over a set of variables in the learned BN, the robot is able to provide information for effects prediction $P(E|O,A)$, feedback in action selection $P(A|O,E)$ or object recognition given its behavioral description $P(O|A,E)$. For example, when we fix a set of perceptual evidence to define an object, and a desired effect,

TABLE I

PREDICTION ERROR FOR EACH VARIABLE IN THE LEARNED BN.

| Tested variable | Prediction error |
|---|---|
| color | 0.1125 |
| size | 0.0725 |
| open_g | 0.1125 |
| end_eff_force | 0.1775 |
| end_eff_mov | 0.025 |
| poi5_mov | 0.025 |
| poi6_mov | 0.055 |
| poi7_mov | 0.05 |
| poi8_mov | 0.05 |

we can obtain a probability distribution of the available actions. The blue object from figure 4, has the highest predicted action probability, for the effect $poi\_obj\_mov$, of $P(lean\_t|obj_{blue}, poi\_obj_{blue}\_mov) = 0.1577$ while the green object which is fixed to the table has zero probability for either manipulation action, due to its nature. Blue object has also a high probability for $poke$ action. These results are coherent with the two affordances (or lack thereof) on these objects: *poke-ability* and *lean-ability*.

## VI. CONCLUSIONS AND PERSPECTIVES

In this paper, we have presented a general approach for learning sensorimotor representations from the unsupervised interaction between perception and action. Bayesian framework captures the relation between the three elements of the affordance definition: effects, objects and actions. Our approach does not rely on *a priori* dependency assumptions between them. It allows the robot to infer the dependencies between the elements while interacting and combining percetual and proprioceptual data. The learned sensorimotor representation along the Bayesian framework allows the robot's motivational system to make predictions about elements in the environment. Moreover, this inferred information can be used for future planning tasks or to add sensor and motor capabilities to the innate repertoire.

We have shown the connection between the three elements of affordance, which allows to represent the learned knowledge in a fused feature frame. We believe that our approach offers a base for further work on the discovery of high level affordances and robot skills.

Currently, we are working on a deep structure learning evaluation using a set of Bayesian-based metrics and information compression metrics to automatically decide the better performance strategy for the affordance learning approach. We are also exploring the learning of innate capabilities of the robot using deep learning techniques to identify salient perceptual features for object representation and salient changes for effect detectors, whereas we made *a priori* hypotheses for them in this paper.

## ACKNOWLEDGMENT

This work has been funded by a DGA (French National Defense Agency) scholarship (ER), and by French Agence Nationale de la Recherche ROBOERGOSUM project under reference ANR-12-CORD-0030.

## REFERENCES

[1] E. Krotkov and R. Bajcsy, "Active vision for reliable ranging: Cooperating focus, stereo, and vergence," *International Journal of computer vision*, vol. 203, no. i, pp. 187–203, 1993.

[2] T. Taniguchi, T. Nagai, T. Nakamura, N. Iwahashi, T. Ogata, and H. Asoh, "Symbol Emergence in Robotics: A Survey," *Advanced Robotics*, vol. To appear, pp. 1–27, 2015.

[3] E. Ugur, Y. Nagai, E. Sahin, and E. Oztop, "Staged Development of Robot Skills: Behavior Formation, Affordance Learning and Imitation with Motionese," *Autonomous Mental Development, IEEE Transactions on*, vol. PP, no. 99, p. 1, 2015.

[4] M. Lungarella, G. Metta, R. Pfeifer, and G. Sandini, "Developmental robotics: a survey," *Connection Science*, vol. 15, no. 4, pp. 151–190, 2003.

[5] M. Lopes, A. Bernardino, and J. Santos-Victor, "A developmental roadmap for task learning by imitation in humanoid robots," *AISB-Third Int. Symp. on Imitation in Animals and Artifacts, Hatfield, Uk*, no. April, pp. 12–14, 2005.

[6] E. Ugur and E. Sahin, "Traversability: A Case Study for Learning and Perceiving Affordances in Robots," *Adaptive Behavior*, vol. 18, no. 3-4, pp. 258–284, 2010.

[7] G. Fritz, L. Paletta, M. Kumar, G. Dorffner, R. Breithaupt, and E. Rome, "Visual learning of affordance based cues," *Biologically Motivated Computer Vision, Proceedings*, vol. 4095, pp. 52–64, 2006.

[8] I. Cos-Aguilera, L. Canamero, and G. Hayes, "Using a SOFM to learn Object Affordances," in *Proceedings of the 5th Workshop on Physical Agents WAF04*, 2004.

[9] S. Griffith, J. Sinapov, M. Miller, and A. Stoytchev, "Toward interactive learning of object categories by a robot: A case study with container and non-container objects," in *2009 IEEE 8th International Conference on Development and Learning, ICDL 2009*, 2009, pp. 1–6.

[10] Y. Demiris and A. Dearden, "From motor babbling to hierarchical learning by imitation: a robot developmental pathway," *International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, pp. 31–37, 2005.

[11] S. Hart, R. A. Grupen, and D. Jensen, "A Relational Representation for Procedural Task Knowledge," in *Proceedings of the 20th National Conference on Artificial Intelligence*, 2005, pp. 1280–1285.

[12] J. Bohg, K. Hausman, B. Sankaran, O. Brock, D. Kragic, S. Schaal, and G. Sukhatme, "Interactive Perception: Leveraging Action in Perception and Perception in Action," *ArXiv e-prints*, 2016. [Online]. Available: arXiv:1604.03670

[13] H. Van Hoof, O. Kroemer, and J. Peters, "Probabilistic interactive segmentation for anthropomorphic robots in cluttered environments," in *IEEE-RAS International Conference on Humanoid Robots*, no. February, 2015, pp. 169–176.

[14] D. Katz, A. Orthey, and O. Brock, "Interactive Perception of Articulated Objects," *Springer Tracts in Advanced Robotics*, vol. 79, pp. 301–315, 2014.

[15] H. van Hoof, O. Kroemer, and J. Peters, "Probabilistic Segmentation and Targeted Exploration of Objects in Cluttered Environments," *IEEE Transactions on Robotics*, pp. 1198–1209, 2014.

[16] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor Segmentation and Support Inference from RGBD Images," in *Proceedings of the 12th European Conference on Computer Vision*, 2012, pp. 746–760.

[17] J. Papon, A. Abramov, M. Schoeler, and F. Worgotter, "Voxel cloud connectivity segmentation - Supervoxels for point clouds," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2027–2034, 2013.

[18] D. Comaniciu, P. Meer, and S. Member, "Mean Shift: A Robust Approach Toward Feature Space Analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 5, pp. 603–619, 2002.

[19] K. Li and M. Q.-H. Meng, "Learn Like Infants: A Strategy for Developmental Learning of Symbolic Skills Using Humanoid Robots," *International Journal of Social Robotics*, pp. 439–450, 2015.

[20] E. Sahin, M. Cakmak, M. R. Dogar, E. Ugur, and G. Ucoluk, "To Afford or Not to Afford: A New Formalization of Affordances Toward Affordance-Based Robot Control," *Adaptive Behavior*, vol. 15, no. 4, pp. 447–472, 2007.

[21] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[22] I. Tsamardinos, L. E. Brown, and C. F. Aliferis, "The max-min hill-climbing Bayesian network structure learning algorithm," *Machine Learning*, vol. 65, no. 1, pp. 31–78, 2006.