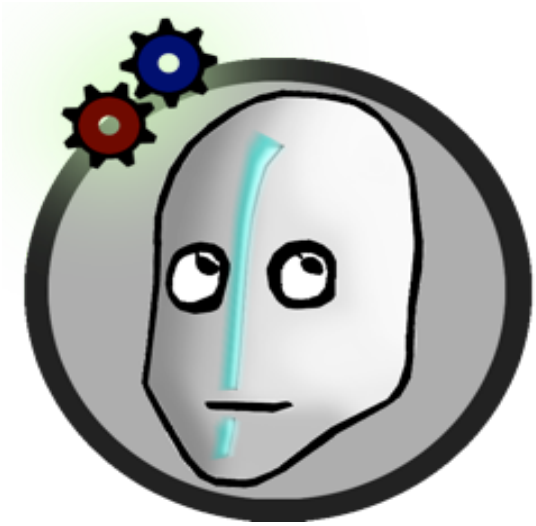

| | |
|---|---|
| **ACRONYME** | **ROBOERGOSUM** |
| **NOM DU PROJET** | ROBOT CONSCIENTS |
| **REFERENCE** | DECISION ANR-12-CORD-0030 |
| | |
| **NUMERO DE LA TACHE** | **T1** |
| **NOM DE LA TACHE** | Situation Awareness and Semantic Scene Interpretation |
| **NUMERO DU RAPPORT** | **D1.1** |
| **TITRE DU RAPPORT** | Sensory-motor Representations and Fusion Processes Discriminating Robot and Environment |
| **PARTENAIRES** | ISIR, LAAS |
| | |
| **DATE** | T0+36 |

## Contents

# 1 Summary

We propose a methodology to build models of objects based on perceptual clues and effects of robot actions on them, which relate to the notion of *affordance*[1]. We employ a Bayesian network that represents with continuous and discrete variables the objects, actions, and effects in the observable environment. We then perform structure learning to identify the most probable Bayesian network that best fits with the observed data. The discovered structure of the Bayesian network allows the robot to discover causal relationships in the environment using statistical data.

# 2 Related work

During recent years, several works have followed a process of exploration-manipulation to understand robot's environment. This process is initialized with a collection of minimal knowledge and innate capabilities which in time is developed through learning following a developmental path.

Deterministic approaches such as [2] and [3], propose learning affordances for predefined categories. Here the relations between perception and action are not discovered exclusively by exploration and relied mainly in supervised classification between predefined classes. In [4] and [5] 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, the object representation of the aforementioned approaches is more related to object recognition rather than object description.

Imitation inspired and probabilistic powered approaches have been explored in [6] and [7]. The latter approach, extracts knowledge trough exploration and reasoning from base knowledge, and uses Bayesian Belief Networks to exclusively associate motor commands to forward models. The former approach, uses relational dependency networks to learn joint probability estimates regarding the effect of sensorimotor features on the predicted desired behaviours. This approach allows a robot to determine which features in the world are relevant to certain motor commands. In both works, the notion of object is hard-wired focusing only in the action-effect association.

# 3 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 [8, 9]. However, the computational cost of inference on these methods rises considerably with the increasing number of objects. Moreover, the relations between nodes come from a priori information from the objects class, which limits their use in self-discovered scenarios.

## 3.1 Over-segmentation approach

Supervoxels are formed by over-segmenting a 3D RGB cloud of points into small regions based on local low-level features (usually geometry and color), reducing the number of nodes which must be considered for segmentation.We implemented a 3D version of the

Voxel Cloud Connectivity Segmentation (VCCS) presented in [10], which takes advantage of 3D geometry provided by RGB-D cameras to generate supervoxels evenly distributed in the observed space, rather than the projected image plane. VCCS uses a seeding methodology based on 3D space and a flow-constrained local iterative clustering which uses color and geometric features. The seeding of supervoxel clusters is done by partitioning 3D space. This ensures that supervoxels are evenly distributed according to the geometry of the scene. The iterative clustering algorithm enforces strict spatial connectivity of occupied voxels. Due to ensuring strict partial connectivity between voxels, this algorithm guarantees that supervoxels cannot flow across boundaries which are disjoint in 3D space even if they are connected in the projected plane.

We define a voxel $i$ inside a voxel cloud $V_r(i) = F_{1..n}$ (from an RGB-D camera) with resolution $r$, as a set of $n$ features defined by vector $F$. First of all, an adjacency graph is constructed for the current voxel-cloud, i.e., for a given voxel, the centers of all $N_a$ adjacent voxels $V_a$ are contained within $\sqrt{3} \times r_v$, where $r_v$ represents the voxel resolution. To ensure that supervoxels borders do not across object boundaries, we use a 26-adjacency.

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 [10]:

$$F = [x, y, z, L, a, b, FPFH_{1..33}] \tag{1}$$

This offers a multi-dimensional pose-invariant representation based on the combination of neighbouring points.

Distance between features is defined by three factors: a normalized spatial distance $D_s$, a normalized color distance $D_c$ and the distance in the FPFH space $D_{HIK}$ calculated by Histogram Intersection Kernel [10]. For each supervoxel, in an outward direction, we calculate the distance from the center of the supervoxel (cluster) to the adjacent voxels. If the distance is the smallest seen, this voxel and its neighbours (in the adjacency graph) become part of the supervoxel.

The result, as is shown in figure 1, 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 [10].

### 3.2 Intrinsic clustering

Figure 1 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 of the number of objects. Regarding the feature representation detailed in Section 3.1, we proposed to use the non-parametric technique described in [11] to find the shape of the object hypotheses based on the set of supervoxels.

Given $n$ data elements (supervoxels) $x_{i..n}$ in a dimensional space $R^d$ and $Li$ the label of the $ith$ element. Initialize $j = 1$ and the initial location of the kernel $H$ as $y_{i,1} = Xi$. Compute $y_{i,j+1}$ according to equation 2 until convergence, i.e., $y = y_{i,c*}$ (see [11] for convergence analysis).

$$y_{j+1} = \frac{\sum_{i=1}^{n} x_i H(||\frac{x-x_i}{h}||^2)}{\sum_{i=1}^{n} H(||\frac{x-x_i}{h}||^2)}, j = 1, 2, ... \tag{2}$$

Assign $z_i = (x_i^s, y_{i,c}^r)$, where $z_i$ represents a filtered version of $x_i$. Superscripts $s$ and $r$ represent the spatial and range components of the vectors. Create a joint domain cluster $\{C_p\}_{p=1...m}$ by grouping together all the $z_i$ which are closer than $h_s$ and $h_r$ in the spatial and range domain, respectively. For each $i = 1, ..., n$, assign $L_i = \{p | z_i \in C_p\}$. Finally, remove regions containing less than $Min_s$ elements.

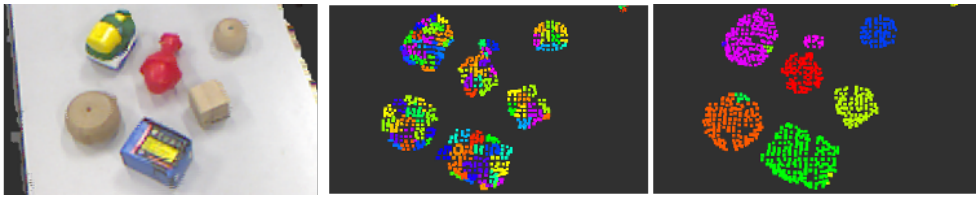

Figure 1: Results from the sensory perception process. RGB-D cloud of points from the real scenario (left); over-segmentation results from point cloud (middle); results from intrinsic clustering (right).

Figure 1 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.

### 3.3 Object hypotheses confirmation

The set of generated hypotheses from Section 3.2 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 object for our sensory-perception task. Each object is represented by its centroid, which offers a point of interaction (*poi*) for the interaction task.

### 3.4 Objects

In this work, we assume that the robot has innate perceptual capabilities that allow it to discretize the environment. This capabilities are related to the segmentation approach. 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 analysing the cloud of points representing the 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 architecture allows for expanding and learning the set of perceptual features.

## 4 Sensorimotor learning

In addition to the perceptual information, object manipulation allows the robot to learn sensorimotor correlations between the sensor inputs fused in the objects descriptions $O$,

robot basic actions $A$ and the salient changes represented by the effects $E$. Starting from built-in actions, the development 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.

### 4.1 Actions

Making an analogy to a newborn rough motor abilities [12], we assume that the robot is built with a set of basic motor capabilities, which we call actions. In this set of basic actions $A = \{a_1, ..., a_n\}$ actions are described relative to actors and their morphology. They are defined with respect to their control variables in joint space:

$$a : \{Q, \dot{Q}, \ddot{Q}\}_\tau \tag{3}$$

where $Q$ are the joint parameters of the robot used in action $a$, and $\tau$ the duration of this action. This implies that, by definition, two actors with completely different motor capabilities and morphologies cannot execute the same action.

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 sensorimotor representation by object manipulation. In a first stage, our proposed set of actions $A$ is composed of 4 actions. Push (*push*) moves relative toward the current position of the end-effector in a constant velocity fashion. Poke (*poke*) behaves in a similar way as *push* but using a constant acceleration dynamic. Open (*open_gripper*) and close gripper (*close_gripper*) are self-explanatory. Due to the general affordance learning goal of our approach, we decided to fix the parameters velocity for *push*, acceleration for *poke*, and force for *close_gripper*. Feedback Information includes the joint and force values of the actuators and the state of the end-effector of the robot.

### 4.2 Effects

An effect is a correlation between an action and a change in the state of the environment, which includes the agent itself. When a robot interacts with an object, it can perceive changes related to the position of the object, proprioceptive values from actuators and feedback from end-effectors. An effect is an important element in our sensorimotor representation 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 (such as object's movement or color change); and proprioceptive-based (e.g. end-effector linear force, distance between gripper's fingers and end-effector's linear movement).

### 4.3 Affordance learning

An affordance is an *acquired* relation between two interacting elements $E$ and $T$, where $E$ is a set of effects and $T$ is a tuple composed of 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 [13]. From an agent's perspective, from now on the robot, the $i_{th}$ affordance is defined as follows:

$$\alpha_i^{\text{agent}} = (e_j, (o_k, a_l)), \text{ for } e_j \in E,\ o_k \in O,\ \text{and } a_l \in A \tag{4}$$

Figure 2 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, $grasp - ability$.

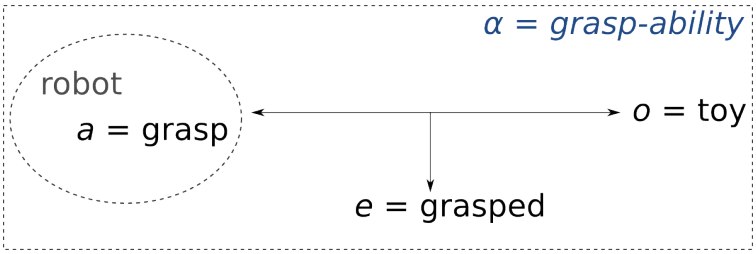


Figure 2: Representation of an affordance relation labelled *grasp-ability*.

We can state our problem as learning the set of relations $\mathcal{A} = \{\alpha_1, ..., \alpha_n\}$ for a set of data extracted from $E$, $O$, and $A$.

## 5 Conclusions

We have presented a methodology for learning sensorimotor representations from the unsupervised interaction between perception and action. Bayesian reasoning captures the relation between the three elements of the affordance definition (4): effects, objects and actions.

Our approach does not rely on a priori dependencies assumptions between objects, actions and effects. It allows the robot to infer the dependencies between these elements while interacting and combining perceptual and proprioceptual data. The learned sensorimotor representations in 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.

## References

[1] J. Gibson, "The theory of affordances," *Perceiving, acting, and knowing: Toward an ecological psychology*, pp. 67–82, 1977.

[2] 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.

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] H. van Hoof, O. Kroemer, and J. Peters, "Probabilistic Segmentation and Targeted Exploration of Objects in Cluttered Environments," *IEEE Transactions on Robotics*, pp. 1–12, 2014.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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. [Online]. Available: http://adb.sagepub.com/content/15/4/447.abstract