Verb Concepts from Affordances

Sinan Kalkan
KOVAN Research Lab, Dept. of Computer Engineering,
Middle East Technical University, Ankara, TURKEY
skalkan@ceng.metu.edu.tr

Nilgün Dağ
KOVAN Research Lab, Dept. of Computer Engineering,
Middle East Technical University, Ankara, TURKEY
nilgundag@ceng.metu.edu.tr

Onur Yürüten
KOVAN Research Lab, Dept. of Computer Engineering,
Middle East Technical University, Ankara, TURKEY
oyuruten@ceng.metu.edu.tr

Anna M. Borghi
EMbodied COgnition Lab, Dept. of Psychology, University
of Bologna and Institute of Cognitive Science and
Technology, National Research Council, Rome, ITALY
annamaria.borghi@unibo.it

Erol Şahin
KOVAN Research Lab, Dept. of Computer Engineering,
Middle East Technical University, Ankara, TURKEY
erol@ceng.metu.edu.tr

In this paper, we investigate how the interactions of a robot with its environment can be used to create concepts that are typically represented by verbs in language. Towards this end, we utilize the notion of affordances to argue that verbs typically refer to the generation of a specific type of effect rather than a specific type of action. Then, we show how a robot can form these concepts through interactions with the environment and how humans can use these concepts to ease their communication with the robots. We demonstrate that iCub, a humanoid robot, can use the concepts, which it has developed, to to *understand* what a human performs, perform multi-step planning for reaching a goal state as well as to specify a goal to the robot using symbolic descriptions.

Introduction

The use of natural language in our interaction with robots remains an elusive target for autonomous robot research. According to the embodied view of intelligence, such a competence requires the robot to link the discrete symbols used in language into meanings that are grounded in the continuous sensory-motor experiences of the robot, infamously named as the *symbol grounding* problem by Harnad (1990).

Although Harnad's approach to intelligence as a *symbol grounding problem* has initiated a great deal of debate, it was well received in the community (Borghi, 2007; Cangelosi & Harnad, 2001; Fischer & Zwaan, 2008; Gallese & Lakoff, 2005). It is now widely accepted that language should be grounded in the sensorimotor experiences of the organism (Cangelosi & Riga, 2006; Cangelosi et al., 2010; Steels, 2003; Glenberg & Kaschak, 2002; Cangelosi, 2010), and that the processing of a word requires the neural circuitry in the brain corresponding to its sensorimotor experience, meaning or simulation (Glenberg et al., 2008; Zwaan & Taylor, 2006). In other words, comprehension of words is likely to involve or require the simulation of the meaning represented by the corresponding concept.

The question that we tackle in this paper can be simply put forward as: *How can a robot ground verbs that we use in our language into its own sensory-motor interactions?* That is, when we command the robot to *push (the table)*[1] it should be able to choose the proper behavior from its own repertoire and apply it. Note that, the behavior chosen for the execution of the verbal command will depend on the subject and that a command such as *push (the cup)* is likely to require a different behavior (Figure 1). Moreover, the very same command will require the use of a behavior executed on the "free" arm of a humanoid robot, who may be holding an object in its other hand. Such a grounding of verbs requires not only the robot to interact with its environment observing the effects it generated, but also the supervision from a human to properly label these effects.

Language, Concepts and Robots

In this section, we summarize the studies and the approaches related to concepts and language in robots and describe the main novelties of our work.

Language in robots

Although comprehension of words should involve the meaning represented by the corresponding concept, the computational efforts in the literature linking language and the sensorimotor data have only focused on mapping a word to a single object or a behavior without much consideration for generalization or conceptualization. Below are summaries of these studies - for reviews on other efforts as well

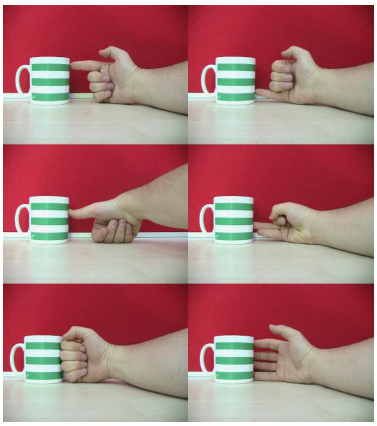


Figure 1. Different behaviors through which one can push an object.

as the importance of perception and action in the development of language, and how and why language should be grounded in action and perception, we refer to Cangelosi (2010); Nehaniv et al. (2007); Christiansen & Kirby (2003); Lyon et al. (2007).

An important tool in linking language and sensori-motor data is artificial neural networks due to its biological plausibility and easy adaptability. Cangelosi (2010) presents a review of their earlier work (all using multi-layer neural networks) on (i) the multi-agent modeling of grounding and language development, using simulated agents that discover labels, or words, for edible and non-edible food while navigating in a limited environment (Cangelosi, 2001), (ii) the transfer of symbol grounding, using one simulated teacher (agent) and one simulated learner (agent) that learn new behaviors based on the symbolic representations of previously learned behaviors (Cangelosi et al., 2006) and (iii) language comprehension in a humanoid robot, where the robot learns to associate words with its behaviors and the objects in the environment. Similarly, in an earlier work, Cangelosi & Parisi (2004) use a neural network for linking nouns to two different objects (a vertical bar and a horizontal bar) and verbs to two different behaviors (pushing and pulling).

Another study on linking language with sensorimotor data (Hashimoto & Masumi, 2007) demonstrates the emergence of symbols (to be linked with language) by interpreting the attractors of a dynamical system (namely, a chaotic neural network) to different symbols and the transitions between the attractors to symbolic manipulation.

For a similar goal, Steels (2007) demonstrates (using a robot and software simulation) the Recruitment Theory of Language, which claims that organisms try different cognitive or motor abilities for communication first and adapt and develop those that lead to successful communication.

Cohen et al. (2005); Krunic et al. (2009); Marocco et al. (2010) have also attributed verbs to individual behaviors without generalization considerations. Other than these, some studies investigate linking behaviors with effects for other purposes: For example, similar to us, Rudolph et al. (2010) proposed relating behaviors to their effects. They suggested that behaviors be represented in terms of their effects. They used their proposal for learning a complex mapping between the hit point and the target point of a thrown ball, and they did not pursue generalization over behaviors or effects nor did they relate their representations to 'verb concepts'. Kozima et al. (2002) have also studied generalization over behaviors based on their effects in the context of imitation; in their *theoretical* proposal, they claimed that a robot should use the equivalence of the effects of behaviors for imitating a human performing a behavior rather than performing geometrical transformations between different embodiments.

Montesano et al. (2008, 2009) proposed a Bayesian network based on an affordance formalization similar to the one used in our paper. In the network, there are nodes for perceptual features (corresponding to the object - *e.g.*, one node for color, one node for shape and another for size), actions (*e.g.*, grasp, tap, touch) and effects (*e.g.*, one node for each of the following: object motion, hand motion and contact of the hand with the object). The model learned the dependencies between the nodes in the network and analyze the learned network in terms of how well it can interpret and imitate an observed effect. Although such an approach has advantages in terms of inferencing, our focus in this paper is different: we are interested in how to generalize over behaviors to be able to represent them as verb concepts allowing efficient "communication" with humans.

As mentioned by Nehaniv et al. (2007), although there exist computational modeling efforts for the emergence of symbols or words for nouns, the emergence of symbolic representations for verbs is still mostly untouched (except for Cangelosi (2010)). Moreover, although highly promising, efforts in grounding verbs (or nouns) mostly do not tackle the issues of generalization over behaviors (or entities) for representing concepts or symbols (e.g., Cangelosi (2001); Cangelosi & Riga (2006); Cangelosi & Parisi (2004)), which is, in fact, the most essential reason for having concepts in a cognitive system.

Theories of Concept

There are three main views on how concepts can be learned or represented (Gabora et al., 2008; Kruschke, 2005; Rosch, 1973; Rouder & Ratcliff, 2006):

- **The Classical, or Rule-based, View**: In this view (see, e.g., (Bruner et al., 1986)), categories are exact with strict boundaries; *i.e.*, an exemplar is either a member of a category or not a member of a category; there is no vagueness involved. The members of the category share common properties (like YELLOW as color and LONG as appearance), and the membership for the category is based on satisfying the common properties of the category, established as rules (like *color of exemplar* = YELLOW $\wedge$ *appearance of exemplar* = LONG).

- **The Prototype-based View**: In this view, the membership for the categories is confidence-based (e.g., (Rosch, 1973)) and the boundaries are not tight. Categories are represented by "prototype" stimuli (the stimuli best representing the category), which are used for judging the membership of

other items. The representation of the prototype is mostly based on statistical regularities, *i.e.*, the frequency distribution of the features, (Ashby & Maddox, 1993). For example, the APPLE concept can be represented by:

$$\text{APPLE} = \left\{ \begin{array}{ll} \textit{color} & \text{50\% RED, 25\% YELLOW or 25\% GREEN} \\ \textit{shape} & \\ \ldots & \end{array} \right.$$

- **The Exemplar-based View**: In this view, concepts are represented by the exemplars of the categories stored in the memory (e.g., (Nosofsky et al., 1992)). An item is classified as a member of a category if it is similar to one of the stored exemplars in that category. For example, the APPLE concept can be represented by:

$$\text{APPLE} = \{ \quad \ldots \}$$

Although the exemplar-based view is in accordance with some experimental results, it falls short in explaining several findings (see (Gabora et al., 2008) for a review and discussion).

Although it is widely believed that the classical view is not adopted by human cognition, there are contradicting evidences about whether humans use prototypes, exemplars or rules for representing concepts (Minda & Smith, 2001; Nosofsky & Zaki, 2002; Leopold et al., 2001). It might be even that for different tasks (such as inferencing or classification), we might be using different types of representations (Johansen & Kruschke, 2005), making a hybrid representation appealing (Rosseel, 2002). Overall, how we represent concepts is still an open issue (Parthemore & Morse, 2010; Gärdenfors, 2004).

Learning concepts is also studied in Machine Learning where efficiency and practicality are the main concerns unlike the theories of concepts in Psychology and the current study, where we are interested in having a developmental conceptualization framework which is biologically plausible (as also discussed in Section "Discussion") and based on enaction. Therefore, we leave an in-depth discussion of the available Machine Learning methods and theories, and refer to Jebara (2004) for a review.

The current study

In this paper, we are interested in how a robot can ground verbs in language. Towards this end, we use the notion of affordances (Gibson, 1986) as formalized by Sahin et al. (2007) to develop a method that can learn to represent and use verb concepts on a humanoid robot platform. Specifically, as the robot interacts with a set of objects using its own repertoire of behaviors, a human observes the effect generated and labels each interaction with a proper verb. The method uses the data collected through such interactions to develop prototypical representations of verbs. Through the use of these prototypes, the robot can be commanded to perform a desired "verb action" on a novel object. The commanding can be provided as (1) a verbal command, such as "push (the cup)", (2) a demonstration, such as the human pushing a box, and asking the robot to mimic what he just did on a cup, (3) goal specification in the prototype space. Moreover, the robot can use these prototypes to make multi-step plans to achieve a goal that is not attainable through a single behavior. Our results have shown that the use of prototypical representations not only reduces the search space for making such plans, but also minimizes the errors in making these plans by paying attention only to the relevant dimensions (as represented in the prototypical representations) in the sensory space.

Affordances and Verbs

The notion of affordance was introduced by Gibson (1986) to propose that organisms perceive the environment in terms of the action possibilities that they offer to them. Gibson argued that when we look at a chair or a cup, our perception does not provide a generic perceptual view of these objects consisting of all of their qualities, but instead informs of the affordances such as *sit-ability* and *lift-ability* that they offer to us.

The notion provided a fresh perspective to the classical theories of perception and has inspired new lines of thinking in a wide range of fields. In an earlier study (Sahin et al., 2007), we formalized this important notion such that it can be utilized to learn and use affordances at different levels of autonomous robot control. In particular, we argued that each interaction episode of an agent with its environment can be represented as an *affordance relation instance* tuple as (Figure 2(a)):

$$(\textit{entity}, \textit{behavior}, \textit{effect}), \qquad (1)$$

where *entity* denotes the environmental relata obtained via perceiving the environment and the self. It encapsulates the perceptual representation of an agent at different complexity levels, ranging from raw sensory data to the features extracted from the environment. However, within the context of this paper, we confine the use of *entity* to a single object. The term *behavior* represents the physical embodiment of the agent's interaction encoding the internal representation that defines a unit of action that can often take parameters for initiation and online control. Within the context of this study, we assume that behaviors are discrete entities. Finally, *effect* is defined as the perceptual change generated in the environment due to the execution of the behavior.

For instance, when a robot applies its *lift* behavior to a *can*, it produces the effect *lifted*, meaning that the can's position, as perceived by the robot, is elevated. Through its interactions with a can, a robot can acquire *relation instances* of the form:

$$(\textit{black-can}, \textit{lift-with-right-hand}, \textit{lifted}),$$

meaning that there exists a potential to generate the effect *lifted* when *lift-with-right-hand* is applied to *black-can*. Note that the term *black-can* is used just as a short-hand label to denote the perceptual representation of the black can by the

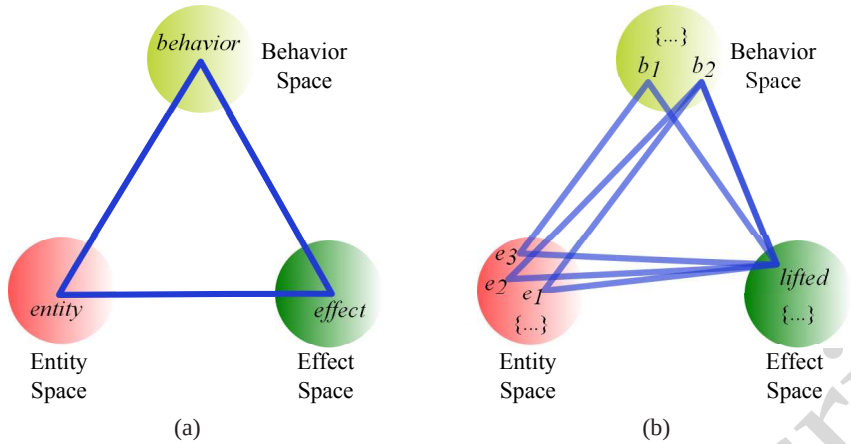


Figure 2. **(a)** An affordance (relation) involves an entity (from the entity space), a behavior (from the behavior) space and an effect (from the effect space) that is produced by applying the behavior on the entity. **(b)** We propose linking verb concepts to generalizations over behaviors based on their effects. In this example, the set of affordance relations that have the *lifted* effect should be linked to the "lift" verb.

interacting agent. Similarly, *lifted* and *lift-with-right-hand* are labels for the related perceptual and proprioceptive representations. For instance, the representation of the black can be a raw feature vector derived from all the sensors of the robot looking at the *black-can* before it attempts to apply its lift behavior.

Arguing that affordances should be relations with predictive abilities, rather than a set of unconnected relation instances, we proposed a learning process that can be applied on this representation. For instance, a robot can achieve the effect *lifted*, by applying the *lift-with-right-hand* behavior on a *black-can*, or a *blue-can*. It can thus learn a relation:

$$(<\text{*-}can>, \textit{lift-with-right-hand}, \textit{lifted}),$$

where <*-*can*> denotes the derived invariants of the can that are relevant for lift-ability. In our previous studies (Uğur et al., 2009; Uğur & Şahin, 2010), we were able to train SVM (Support Vector Machine) classifiers to implement prediction modules such as <*-*can*> for each behavior, successfully. In these studies, the effects were grouped into a number of discrete effect categories

Affordances and Language

The link between the notion of affordances and language comprehension has already been pointed out in Psychology (Borghi & Riggio, 2009; Borghi, 2012). The indexical hypothesis by Glenberg & Robertson (2000) explains how this may happen. According to the hypothesis, words and sentences are linked to objects in the world, their referents, or to analogical representations as pictures or perceptual symbols (Barsalou, 1999). For example, the word *handle* refers to its referent, a handle, or to an analogical representation of the handle. Thus words that refer to objects would evoke firstly perceptual information relative to such objects. Given the close relationship between perceptual and motor processes, words should also evoke motor information. Indeed, depending on their perceptual features, objects can activate affordances (Gibson, 1986). For instance, different kinds of handle may afford different actions: some can be turned, some pushed to open a door. From this view comes the idea that activation is more tied to the affordances elicited by objects than to the words representing the objects. Object affordances would influence not only the understanding of words but also the understanding of more complex linguistic structures such as sentences.

Although the relationship between words, concepts and affordances has been pointed out by others, the problem of how such a link exists in organisms and how it can be created in robots has not been completely tackled yet. In this article, we argue that verbs that are provided by a human observing the physical interactions of the robot with objects can be used to bridge the concepts represented by these verbs into sensorimotor interactions of the robot.

Within the context of this paper, we assume that verbs, that are used to command a robot, mostly specify the accomplishment of a desired goal with no regard on the means of how it is achieved. For instance, when we command a robot to *lift* (a box), we expect him to pick the proper behavior to vertically elevate the box. As illustrated in Figure 1, the command should invoke different behaviors on the robot as determined by the properties of the box (such as size) or the state the robot (such as the robot already holding a cup in one of its hands). Such an ability relieves the human from being aware of the robot's sensorimotor capabilities and requires the robot to flexibly respond to verbal commands based on its prior interaction with the objects.

Verbs: behavior or effect categories

It is tempting to associate the concept of a verb with a category that covers all the interactions that are generated by the execution of a particular behavior. If we want the robot to lift a particular object[2], the verb "lift" can trigger the lift behavior of the robot to accomplish our goal. For instance, it

might be suggested that the concept of lifting should cover:

$$(\left\{ \begin{array}{c} \cdot \\ \cdot \\ \cdot \end{array} \right\}, lift, \left\{ \begin{array}{c} \cdot \\ \cdot \\ \cdot \end{array} \right\}). \quad (2)$$

However, such an association provides a limited coverage for all the meanings that the verb "lift" should convey. First, the robot can probably lift an object with different behaviors, such as *lift-with-right-arm* and *lift-with-left-arm* (for example, Figure 1 shows six different behaviors that can be used by humans to push an object towards left). Second, the execution of the particular behavior may fail on some objects, e.g., heavy or slippery objects. Third, in certain cases, a seemingly contradictory behavior such as pressing, may also lift an object that is placed on a lever to accomplish lifting.

The criticisms that are stated above indicate that the representation of a verb concept by a particular behavioral category implicitly includes the "manner" information by specifying the exact type of behavior that is being asked for. An alternative, which we take in this article, is to associate verbs with effect categories as:

$$(<any\text{-}entity>, <any\text{-}behavior>, lifted). \quad (3)$$

In other words, we propose linking the verb "lift" to the set of behaviors that have the *lifted* effect (see Figure 2(b)).

Experimental Framework

We used the iCub humanoid robot (Metta et al., 2008), a fully open-source platform designed for cognitive and developmental robotics research. The robot, built in the form of 4 year old child, has 53-DOF in its body and equipped with 7-DOF arms and 9-DOF hands making it possible to develop human-like simple object manipulation behaviors for interacting with objects put on a table.

The robot used a Kinect RGB-D camera (Figure 3) fixated on the side of the robot to perceive the objects on the table. The camera captured the depth of scenes with a resolution of 640×480, providing a cloud of 3D points with the corresponding RGB data.

Behaviors

We used a repertoire of six manipulation behaviors for interacting with the objects, similar to the ones used by Bergquist et al. (2009); Metta & Fitzpatrick (2003). These behaviors, denoted as $b_0, .., b_5$, are: *push-left*, *push-right*, *push-forward*, *pull*, *top-grasp* and *side-grasp* behaviors[3]. The *top-grasp* and *side-grasp* behaviors are approach the object from the top, or from the left or right (depending on the relative position of the object) and fingers close upon touch.

Perceptual features

The object in the depth image captured by the Kinect device is segmented from the tabletop by assuming that the workspace is planar and placed parallel to the ground. The

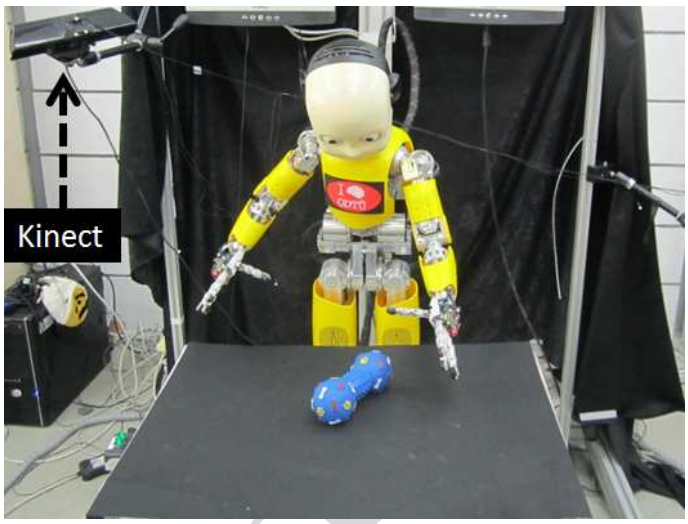


Figure 3. iCub interacting with an object on the table.

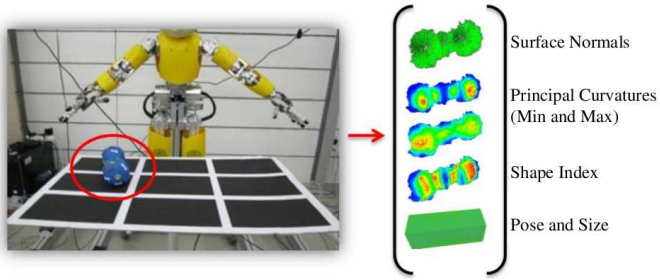


Figure 4. The elements of perception extracted within our system.

following features were then extracted from the point cloud corresponding to the object:

- *Surface features:* surface normals (azimuth and zenith angles), principal curvatures, and shape index as represented with 20-bin histograms, using curvature and normal estimation methods provided by an open-source Point Cloud Library - PCL (Rusu & Cousins, 2011).
- *Spatial features:* bounding box center, orientation, and dimensions (along *x*, *y*, *z*).
- *Object Presence:* a binary feature for whether an object exists on top of the table or not. This information is especially useful when an object disappears after an interaction.

The features extracted from the objects before the execution of a behavior are called the *initial features* whereas the features extracted after the behavior are called the *final features*. The difference between the final and the initial features are used as the *effect features*. These initial and effect features correspond to the *entity* and the *effect* in the affordance formalization in Equation 1.

Learning Affordance Relations

In the experiments, the robot interacted with a set of 35 objects of different sizes and shapes as shown in Figure 5. In total, 413 different interactions were recorded, that consisted

Figure 5. The objects interacted by the robot for learning.

of multiple interactions with the objects placed at different positions and in different orientations, in order to capture the variability.

During these interactions, the initial and final features of the objects were recorded, and the effects generated on the objects were labeled by a human. Specifically, each interaction episode is encoded as a relation between an object $o_j \in O$, a behavior $b_i \in \mathcal{B}$ and an effect f as:

$$(e_{o_i}, b_j, f_{o_i}^{b_j}), \tag{4}$$

where e_{o_i} is the initial perceptual representation of the object o_i; $b_j \in \mathcal{B}$ is a behavior from the set of behaviors $\mathcal{B}$; and $f_{o_i}^{b_j}$ is the representation of the effect. The effect $f_{o_i}^{b_j}$ is defined as the difference observed in the perceptual representation of object e_{o_i} as a result of the interaction as:

$$f_{o_i}^{b_j} = e_{o_i}^{b_j} - e_{o_i}. \tag{5}$$

Then, each interaction is labeled by a human based on the effect generated using a set of verbs (*i.e.*, effect labels) $E \in \mathcal{E}$ where $\mathcal{E}$ included `no-effect`, `moved-left`, `moved-right`, `moved-forward`, `pulled`, `grasped`,`knocked`, and `disappeared`. For example, if the robot applies a *push-right* behavior on an object, leading to a measurable displacement towards the right, the user verbally provides "moved right" to the robot.

Figure 6 depicts the categories formed in the effect space as a result of the effect labeling. For instance, when the robot applied the *push-left* behavior on cubes and cylinders, the objects `moved-left`. However, the application of the very same behavior on the balls, caused the objects to `disappeared`, since they rolled away and became invisible. It can also be seen that the same `disappeared` effect can be generated on balls through the application of `push-*` (any type of push) behaviors.

The overall process of learning affordances from the affordance relation instances is sketched in Figure 7. Specifically,

Table 1
The average, maximum and minimum prediction accuracies of SVMs for each behavior obtained through 5-fold cross validation.

Behavior	Average Accuracy	Maximum Accuracy	Minimum Accuracy
side-grasp	100%	100%	100%
top-grasp	90%	100%	85%
push-left	92%	100%	83%
push-right	96%	100%	85%
push-forward	100%	100%	100%
pull	96%	100%	86%

for each behavior b_i, the mapping $\mathcal{M}_{b_i} : e_{o_j} \rightarrow E_{o_j}^{b_i}$ from the initial representation of the objects (*i.e.*, e_{o_j}) to the effect clusters $E_{o_j}^{b_i}$ is learned by a Support Vector Machine (SVM) classifier. These SVMs enable the robot to predict the effect category ($E_{o_l}^{b_k}$) that it can generate by applying a behavior b_k on a novel object o_l. In our experiments, the SVM classifiers for each behavior were trained with 5-fold cross validation reaching average accuracy values above 90% (as can be seen in Table 1).

We would like to note that these SVMs effectively provide an affordance-based perception view of the object, by predicting what the robot can do (such as move-right, knock, disappear etc.) with them, i.e. what they afford.

Verb Conceptualization

In this section, we describe (i) our verb conceptualization based on effect prototypes and two alternative methods for verb conceptualization, and (ii) how verb concepts can be used for various human-robot interaction problems.

I - Verb Conceptualization Using Effect Prototypes - C_{EP}

In this section, we describe how we derive the condensed prototype representation f_{pro} of the effects $\{f\}$ in an effect cluster $E \in \mathcal{E}$ (Figure 6). We call this condensed representation the *effect prototype* and claim that they correspond to *concepts* represented by verbs.

Figure 8 depicts a summarized version of the distribution of effect features for different effect categories. Examining the distribution of change in each feature element, *i.e.*, ${}_i f$ (where ${}_i f$ is the i^{th} element of the n-dimensional feature vector f), we observe four different characteristics: feature elements that (i) increase consistently, (ii) decrease consistently, (iii) remain constant or (iv) change in an unpredictable way. Therefore, we find it suitable to represent an effect prototype using labels '+', '-', '0', '*', corresponding to increase, decrease, no-change and unpredictable-change in the feature element, respectively. In addition to these labels, we also include the mean and the variance of the changes in the representation to quantify the amount of the changes.

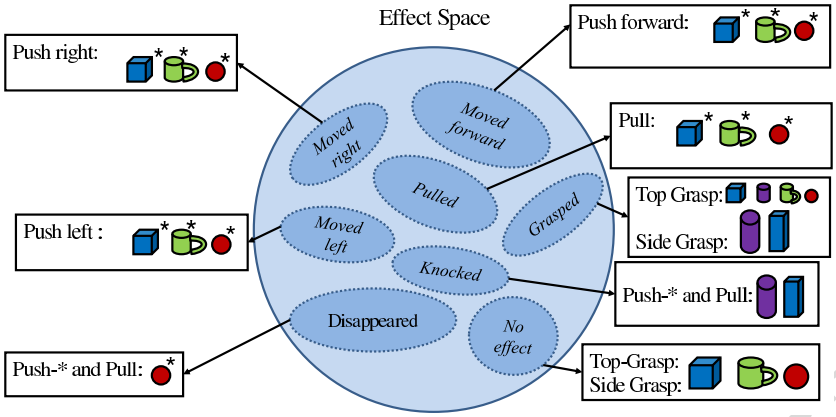


Figure 6. Labeled clusters in the effect space. '*' represent all instances of the corresponding object category.

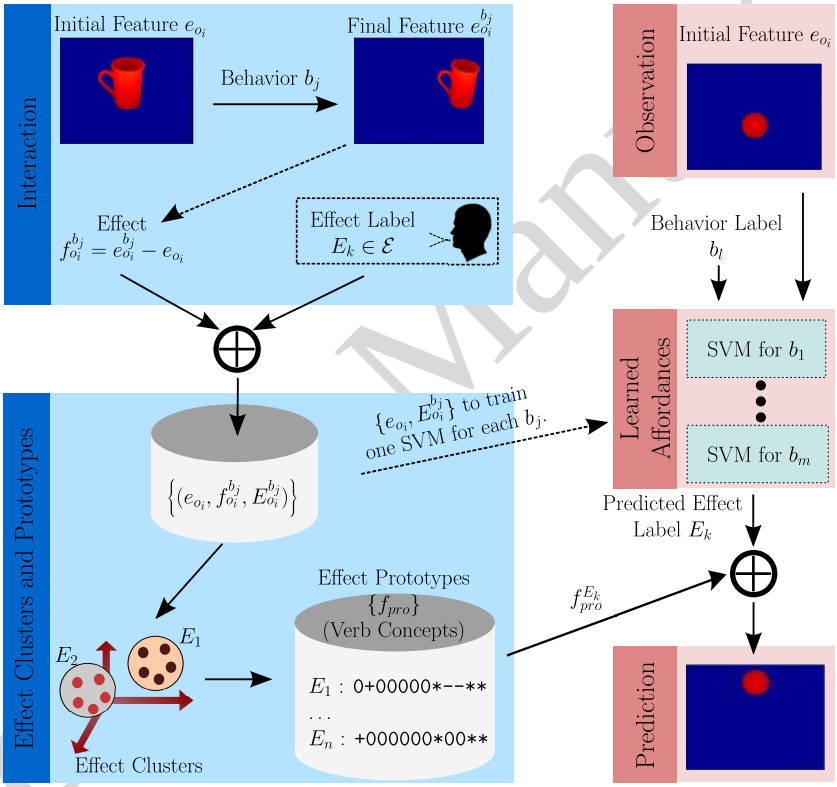


Figure 7. Clusters in the effect space are used for training an SVM, which allows the effect label to be predicted from a behavior on a novel object.

As a result, we define an *effect prototype* as a string consisting of labels '+', '-', '0', '*', called "prototype labels" in the rest of the article, together with two vectors corresponding to the mean and the variance of the observed changes. In order to assign prototype labels to the feature elements, we use unsupervised clustering (namely, Robust Growing Neural Gas (Qin & Suganthan, 2004)) in the space of mean and variance of the changes (summarized in Algorithm 1). The prototypes derived from our experiments are shown in Table 2. For the sake of clarity, we will abbreviate these prototypes as a combination of s^k denoting k consecutive occurrences of the symbol s (which can be '+', '-', '0' or '*').

In order to compare two effect prototypes or an effect prototype with an effect instance, we define a similarity metric using the Mahalanobis distance (Mahalanobis, 1936). This modified version of Mahalanobis distance between two effect clusters (or between an effect prototype and an effect instance - Equation 8) is calculated by taking the mean μ_{E_i} of first effect cluster E_i and using the second effect cluster's E_j mean μ_{E_j} and variance σ_{E_j}:

$$d_{EP}(E_i, E_j) = \sqrt{\left(\mu_{E_i} - f^{+,-,0}_{pro,E_j}\right)^T S_j^{-1} \left(\mu_{E_i} - f^{+,-,0}_{pro,E_j}\right)}, \quad (8)$$

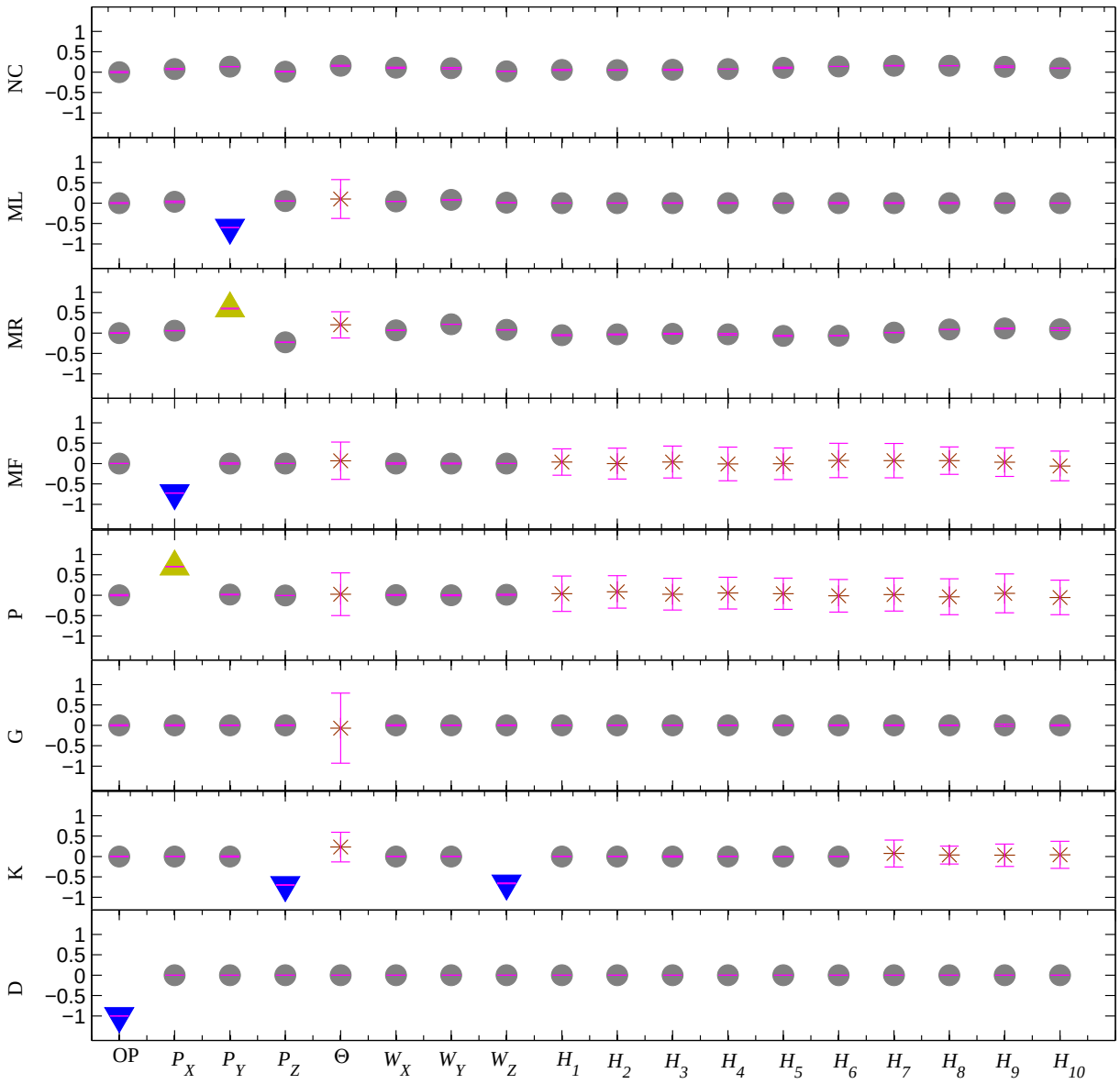


Figure 8. An illustration of how feature dimensions change in effect clusters. The types of changes are obtained by unsupervised clustering (using RGNG - see the text) in the space of mean and variance of the changes. The shapes (circle, triangle and star) in this plot correspond to the mean values of the changes, while the error bars correspond to their variance (in the case of circles and triangles, the error bar looks like a single line due to small variance). From unsupervised clustering of the changes, we get four change types (clusters): consistently increasing (upwards triangle), consistently decreasing (downwards triangle), consistently not changing (circle) and inconsistenly changing (star). The abbreviations on the y-axis stand for some of our effects (NC: *"no change"*, ML: *"moved-left"*, MR: *"moved-right"*, MF: *"moved-forward"*, P: *"pulled"*, G: *"grasped"*, K: *"knocked"* and D: *"disappeared"*) and the abbreviations on the horizontal axis stand for the feature elements (OP: *Object Presence*, P_X: *position-x*, P_Y: *position-y*, P_Z: *position-z*, Θ: *Orientation*, W_X, W_Y, W_Z: Size along *x*, *y* and *z* and $H_1 ... H_{10}$: *Shape Histograms* - **only a subset of the shape histograms are provided for the sake of space, see Table 2 for a complete listing**).

where S_j is the covariance matrix of the second effect cluster E_j. In accordance with (Verguts et al., 2004) who claim that (i) non-existing features and (ii) dissimilar features are not used in computing similarity between categories, in computing the Mahalanobis distance, the dimensions denoted by a '*' in the prototype strings are disregarded (denoted by $f^{+,-,0}_{pro,E_i}$ for the effect prototype f_{pro,E_i} of an effect cluster E_i), since these correspond to an unpredictable/inconsistent change in the feature elements.

Table 2
Effect prototype strings that are extracted using the effect-prototype conceptualization (C_{EP}) introduced in Section "Verb Conceptualization". Note that each feature element has an associated mean and variance of the change (not shown here).

Eff. Cat. Name	Δ Azimuth Histograms	Δ Zenith Histograms	Δ Curvature Histograms	Δ Shape Index Histograms	Δ Position (x-y-z)	Δ Orient.	Δ Size (x-y-z)	Δ Object Presence
NC	*000000000 0000000000	0000000000 0000000*00	0000000000 00*0000000	0000000000 000000*000	000	0	000	0
MR	********** ******0*0*	********** ********000	0000000000 00*********	*****0**** **0*0*0***	*+*	*	***	0
ML	******0000 00***0****	0**0**0000 **0000*000	0000000000 0*****0000	00*0**0*0* **0000*00*	0-0	*	000	0
MF	********** **********	******000* **000**000	0000000000 00****0***	***0*0**** **00000***	-00	*	000	0
P	********** **********	******000* **000**000	0000000000 00****0***	***0*0**** **00000***	+00	*	000	0
K	*0***00000 000000**0*	*000000000 0000000000	0000000000 000000000*	0000000000 0000000000	00-	*	00-	0
G	0000000000 0000000000	0000000000 0000000000	0000000000 0000000000	0000000000 0000000000	000	*	000	0
D	0000000000 0000000000	0000000000 0000000000	0000000000 0000000000	0000000000 0000000000	000	0	000	-

Algorithm 1 Derivation of Effect Prototypes - for C_{EP}.

Given: Interactions with the environment to collect a set of effects $\left\{f_{o_i}^{b_j} \mid \forall b_j \in \mathcal{B}, \forall o_i \in O\right\}$.
Output: Effect prototypes f_{pro} (*i.e.*, C_{EP}) for each effect category.

- Assign a label $E \in \mathcal{E}$ to each effect.
for all E in the set of effect clusters $\mathcal{E}$ **do**
- Compute the mean ${}_i\mu_E$ of the change in each feature element i:

$$ {}_i\mu_E = \frac{1}{N} \sum_{f \in E} {}_i f, \tag{6}$$

where N is the cardinality of the set $\{f \in E\}$.
- Compute the variance ${}_i\sigma_E$ of the change in each feature element i:

$$ {}_i\sigma_E = \frac{1}{N} \sum_{f \in E} ({}_i f - {}_i \mu_E)^2 . \tag{7}$$

end for
- Apply Robust Neural Growing Gas (RGNG) algorithm (Qin & Suganthan, 2004) in the space of $\mu \times \sigma$.
- Manually assign the labels '+', '-', '0' and '*' to the four clusters that emerge in the previous step.

II - Verb Conceptualization Using Naive Prototypes - C_{NP}

In order to evaluate our effect-prototype-based verb concepts (C_{EP}), we introduce another prototype representation of verbs that do not utilize the string representation (*i.e.*, '+', '-', '*', '0'). This amounts to pure Mahalanobis distance which considers all dimensions in a feature:

$$ d_{NP}(E_i, E_j) = \sqrt{\left(\mu_{E_i} - f_{pro,E_j}\right)^T S_j^{-1} \left(\mu_{E_i} - f_{pro,E_j}\right)}. \tag{9}$$

III - Verb Conceptualization Using Exemplars - C_{Ex}

For better evaluation, we also introduce conceptualization of verbs using the exemplars in the categories. In this case, checking the membership of an instance requires comparing that instance with all the members of a category and picking up the category that has the minimum distance. Itemwise comparison is achieved using Euclidean distance:

$$ d_{Ex}(f_{new}, E_i) = \min_{f \epsilon E_i} \sqrt{\sum_{n=1}^{N} ({}_i f_{new} - {}_i f)^2}, \tag{10}$$

where f_{new} is the new effect instance; E_i is the effect cluster f_{new} is compared against; and, N is the number of dimensions in a feature.

Understanding an interaction in terms of verbs

An important problem in human-robot interaction is the correspondence between the different embodiments (Alissandrakis et al., 2003), which requires, e.g., matching the different body parts of the human to the parts of the robot. A practical way around the correspondence problem is to interpret interactions based on their effects using the symbolic space of effect prototypes. In this study, matching an ob-

served interaction (effect) with another effect or prototype (C_{EP}, C_{NP} or C_{Ex}) is achieved using the distance functions d_{EP}, d_{NP} and d_{Ex} respectively provided in Equations 8, 9 and 10, as outlined in Algorithm 2.

Goal specification through demonstration and multi-step planning

A natural way to command a robot is to specify our goal through demonstration, a form of non-verbal communication that humans use with babies, with people that we do not have a common language, or with people that we have to communicate in loud environments. We term this form of communication as *goal specification through demonstration* in general. Within the context of this study, we would like a human to demonstrate a desired goal, by demonstrating it in front of the robot and ask him to "do what I just did". In this study, we can achieve this by using verb concepts, which provide abstraction over the behaviors, eliminating the need to recognize individual behaviors and to handle the correspondence problem (Alissandrakis et al., 2003).

Our method for "do what I just did" (see also Algorithm 3) relies on (i) predicting the outcome of each behavior, (ii) comparing the predictions with the desired observed effect (*i.e.*, what the human has demonstrated) and (iii) repeating step-(i) for each prediction produced in step-(ii). For comparing the predictions with the desired effect, we will use and compare the distance functions d_{EP}, d_{NP} and d_{Ex} respectively provided in Equations 8, 9 and 10.

Commanding with verbs or symbols

The SVMs allow the robot to predict the category of the effect that it can generate on a novel object after executing a certain behavior. This allows the robot to respond to verb commands, such as push-right (the object on the table), by feeding the objects perceptual representation to all the SVMs and checking whether the specified goal (via giving the verbal command) matches with the predicted effects of any of the behaviors as outlined in Algorithm 4. Moreover, the robot can be specified a goal in terms of '+', '-', '0' or '*' symbols, and satisfy such a goal by finding the behavior yielding the closest effect to the specified goal (using the distance functions Equations 8, 9 and 10).

Note that the application of more than one behavior may be predicted to generate the desired effect specified by the commanding verb. The set of these behaviors provides the robot with a *flexibility* that can be useful in cases of failure or in making multi-step plans (as outlined in Algorithm 3) subject to other constraints.

Results

In this section, we first demonstrate and evaluate the three different verb conceptualization methods outlined in the previous section. Having verbs or verb concepts should enable an organism (1) to understand, in his own sensory-motor and symbolic representations, the observed behavior of another organism, and (2) to achieve goals specified in his own symbolic representations which are grounded in his own sensory-motor system. We demonstrate and evaluate both aspects on the iCub platform.

Algorithm 2 Understanding an observed effect/behavior.

Given: Observation of an entity e_{o_i} and a behavior applied on e_{o_i}, leading to the effect $f^j_{o_i}$. Note that this behavior may not be in the repertoire of the robot.
Output: Determine the verb concept E^* (*i.e.*, the effect category) that best describes the observed interaction.

- Take E^* (the best matching effect category) as the interpretation of the observed effect:

$$E^* = \arg\min_{E \in \mathcal{E}} d_C(f^j_{o_i}, f_{pro,E}), \tag{11}$$

where $d_C(.,.)$ is either $d_{EP}()$, $d_{NP}()$ or $d_{Ex}()$ respectively defined in Equations 8, 9 and 10; and, $f_{pro,E}$ is the prototype of the effect category E. If required, a threshold on $d_C(f^j_{o_i}, f_{pro,E})$ can be set as a criteria to determine whether the observed effect is unknown to the robot.
- (Optional) Given a novel entity e_{o_k}, find the behavior b^* (among the behavior repertoire of the robot) that produces an effect in the effect cluster represented by the effect prototype f_{pro,E^*}:

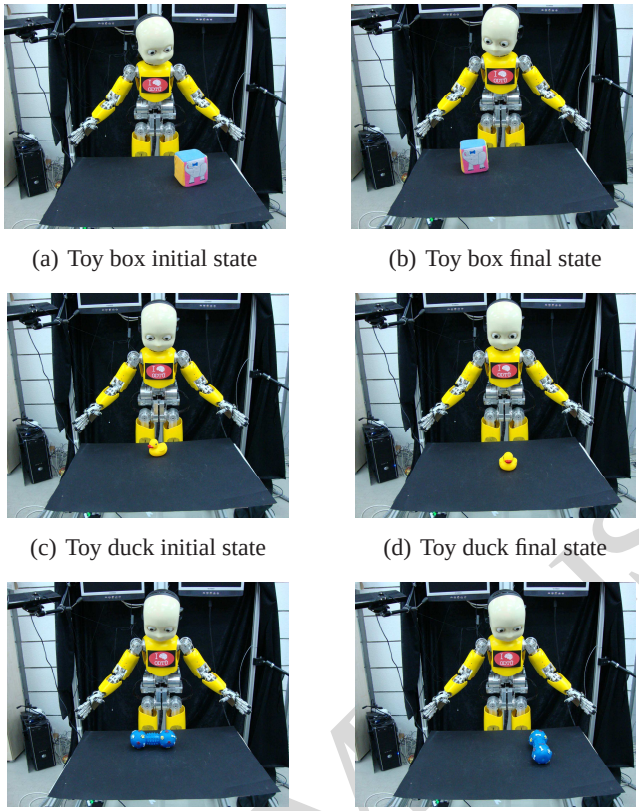
$$b^* = \arg\max_{b} d_C(\text{SVM}(e_{o_k}, b), f_{pro,E^*}), \tag{12}$$

where $d_C(.,.)$ is either $d_{EP}()$, $d_{NP}()$ or $d_{Ex}()$ respectively defined in Equations 8, 9 and 10; and f_{pro,E^*} is the prototype of the effect category E^*. If required, a threshold on $d(\text{SVM}(e_{o_k}, b), f_{pro,E^*})$ can be set as a criteria to determine whether the observed behavior cannot be replicated on the novel object e_{o_k}.

Verb concepts for goal emulation and multi-step planning

In Figure 9, some novel interactions (leading to novel effect instances) are shown. For these instances, the robot can find the best interpretation by matching them against the verb concepts that it has formed using the distances defined in the previous section.

For these effect instances, we compare our prototype-based representation (C_{EP}) with the naive prototype representation (C_{NP}) and the exemplar-based representation (C_{Ex}), as shown in Table 3. We see that our prototype-based representation can find the correct category whereas the exemplar-based conceptualization and the naive prototype-based conceptualization fail to find the correct category in some cases. C_{Ex} especially fails because the observed instances are closest to the *disappeared* effect category since all dimensions in this category are zero. C_{NP} performs better



(a) Toy box initial state (b) Toy box final state

(c) Toy duck initial state (d) Toy duck final state

(e) Toy bone initial state (f) Toy bone final state

Figure 9. Some novel interactions with novel objects. The effect is simply the difference of final and initial states of the given object. The novel effect predictions with different distance metrics for these instances are listed in Table 3.

than C_{Ex}; however, we see that inconsistent dimensions that are not excluded by C_{NP} in distance calculations may lead to wrong effect categories.

Another advantage of the prototype-based representation is that iCub can symbolically describe what it has seen. In Figure 9, iCub is shown two different interactions. Observing the effects, iCub finds the effect prototype (Figure 10) that best describes the observed behavior using Algorithm 2. The matching effect prototype is the symbolic representation (*i.e.*, the verb concept) of the observed behavior and this symbolic representation is grounded in iCub's sensori-motor experiences. Having the sensori-motor grounding of the effect prototype, iCub is asked to produce the same effect (Figure 10). Note that with the set of behaviors iCub is equipped with, there may be more than one way to achieve the goal, and iCub chooses the one with highest prediction accuracy, as described in Algorithm 2.

In a scenario requiring multi-step plans, we compare the prototype with the exemplar-based and naive prototype conceptualization. The multi-step planning results are provided in Figure 11. We see that, using the verb concepts presented in Table 2, iCub can successfully find a sequence of effect prototypes leading to the target state. From these prototypes, iCub can choose the best behaviors that can generate those effects. In Figures 13 and 14, we provide the planning results when naive prototypes (C_{NP} - Equation 9) or exemplars (C_{Ex} - Equation 10) are used for conceptualization. We see that, in these cases, the planner could not produce proper behavior sequences to achieve the given goals in limited steps (the distance threshold was constant throughout the experiments). The multi-step planning is sketched in Algorithm 3.

Verb concepts and goal specification

As if specifying a goal for iCub with a verb (like "push left the object"), we give iCub the goal with his own symbolic representations (Algorithm 4). Since they are grounded in iCub's sensori-motor experiences, iCub can find the behavior that satisfies the requested goal (shown in Figure 15).

Discussion

In this article, we have taken an ecological, embodied and grounded approach to verb conceptualization. We have proposed novel methodologies for linking the notion of affordances with concepts that correspond to verbs in language. To this end, a humanoid robot, iCub exercised its behavior repertoire on the objects available in the environment for the purpose of discovering the affordances of the objects.

The learned affordances allow the derivation of novel condensed representations of behaviors' effects, which we called

Table 3
Evaluation of the different conceptualization methods (i.e., C_{EP}, C_{NP}, C_{Ex}) for the novel interactions in Figure 9. The table lists the distances between the observed effect and the existing verb concepts. The verb concept with the smallest distance is the corresponding interpretation of the corresponding conceptualization method (i.e., one of C_{EP}, C_{NP}, C_{Ex}). The correct predictions are in bold, whereas false predictions are underlined.

Inter.	Concept.	No Change	Moved Right	Moved Left	Moved Forward	Pulled	Knocked	Grasped	Disappeared
Figure 9(b)	C_{EP}	390.81	**146.24**	372.24	389.21	215.56	215.50	392.11	410.31
	C_{NP}	392.16	**182.13**	386.92	416.43	241.06	219.28	395.04	410.31
	C_{Ex}	237.01	236.89	237.42	237.24	237.42	237.42	237.25	236.84
Figure 9(d)	C_{EP}	731.36	494.18	416.42	**340.71**	393.76	358.06	738.04	790.41
	C_{NP}	732.98	497.02	417.18	426.71	423.17	428.06	741.11	790.41
	C_{Ex}	789.45	789.08	789.83	789.49	789.83	789.83	789.54	788.84
Figure 9(f)	d_{EP}	925.41	577.51	**267.45**	328.75	354.85	354.74	928.16	947.51
	C_{NP}	929.37	580.26	**291.77**	369.75	373.37	359.88	929.94	947.51
	C_{Ex}	946.74	946.42	947.03	946.66	947.03	947.03	947.01	946.21

Algorithm 3 Multi-step planning algorithm

Given: e_{start} and e_{goal}.
Output: P, a plan, which is a sequence of behaviors leading to e_{goal} from e_{start}.

- Initialize: $e_{current} \leftarrow e_{start}$.
for all *level* $= 1 : N_{level}$ **do**
- Update the remaining effect: $f_{current} \leftarrow e_{goal} - e_{current}$.
- Find the verb concept that is closest to e_{goal}:

$$E^* = \arg\min_{E \in \mathcal{E}} d_C(f_{current}, E). \quad // d_C(): d_{EP}(), d_{NP}() \text{ or } d_{Ex}() \tag{13}$$

where $d_C(.,.)$ is either $d_{EP}()$, $d_{NP}()$ or $d_{Ex}()$ respectively defined in Equations 8, 9 and 10.
- Find the behavior that takes us closer to e_{goal}. This behavior is the one that best produces an effect corresponding to the verb concept E^*:

$$b^* = \arg\min_{b \in \mathcal{B}} d_C(SVM(e_{current}, b), E^*). \tag{14}$$

where $d_C(.,.)$ is either $d_{EP}()$, $d_{NP}()$ or $d_{Ex}()$ respectively defined in Equations 8, 9 and 10.
- Update the plan by adding the new behavior: $P \leftarrow P + b^*$.
- Update the current state of the object using the predicted verb concept E^*:
if C_{EP} **then**
$e_{current} \leftarrow e_{current} + f^{+,-,0}_{proto,E^*}$
else
// C_{NP} *or* C_{Ex}
$e_{current} \leftarrow e_{current} + \mu_{E^*}$
end if
end for

Algorithm 4 Satisfying a given symbolic goal specification.

Given: f_{goal}, which is a rough description of what should change in what direction (marked with '+' and '-'). If required, the user can also specify what should not change (with a '0'). The other elements are marked as '*'.
Output: Find b^* (among the behavior repertoire of the robot) that satisfies the goal f_{goal}.

- Take f^*_{pro} (the best matching effect prototype) as the interpretation of the goal:

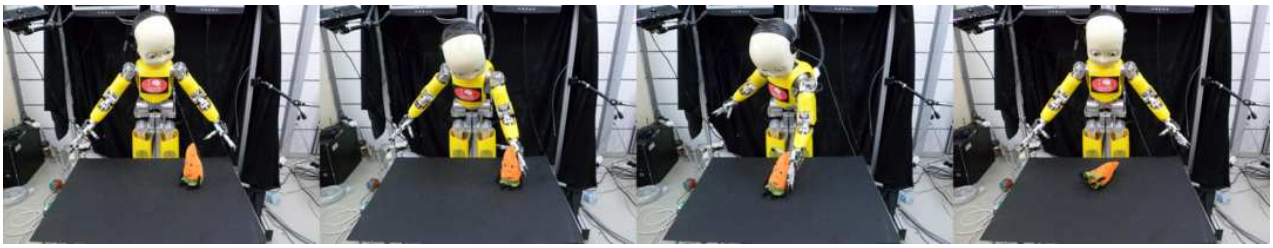
$$f^*_{pro} = \arg\min_{f_{pro}} d_{EP}(f_{\text{goal}}, f_{pro}), \tag{15}$$

where $d_{EP}(.,.)$ is the Mahalanobis distance in Equation 8. If required, a threshold on $d_{EP}(f_{\text{goal}}, f_{pro})$ can be set as a criteria to determine whether the goal cannot be satisfied by the robot.
- Given a novel entity e_{o_k}, find the behavior b^* (among the behavior repertoire of the robot) that produces an effect in the effect cluster represented by the effect prototype f^*_{pro}:

$$b^* = \arg\max_{b} d_{EP}(\text{SVM}(e_{o_k}, b), f^*_{pro}), \tag{16}$$

where $d_{EP}(.,.)$ is the Mahalanobis distance in Equation 8. If required, a threshold on $d_{EP}(\text{SVM}(e_{o_k}, b), f^*_{pro})$ can be set as a criteria to determine whether the goal cannot be achieved on the novel object e_{o_k}.

effect prototypes. We proposed that effect prototypes correspond to verb concepts. We demonstrated that, with these concepts, the robot can generalize/abstract over its behaviors, and represent the behaviors (and what they are useful for) using symbols, which then allow the robot to interpret its own or others' interactions with the environment. These concepts can easily be linked to words (*i.e.*, verbs like "push", "lift", etc.) through which the robot can interact with humans



iCub computes the category of the effect in Figures 9(a)-9(b) as "`moved right`"

iCub chooses and applies *push-right* on the object

iCub computes the category of the effect in Figures 9(e)-9(f) as "`moved left`"

iCub chooses and applies *push-left* on the object left

Figure 10. "Do what I just did" Demonstration. **First row:** iCub interprets the interaction in Figures 9(a)-9(b) as an instance of "`moved-right`" *verb concept*, i.e., which only has the change in *y* position as consistently increasing, more specifically: $*^{16}[0*]^2\ *^{17}\ 0^{15}\ *^{13}\ 0\ *^6\ [0*]^3\ *^3 + *^5\ 0$ (For the sake of space, we denote k consecutive occurrences of a symbol s with s^k). Then, iCub is asked to create the same effect on a novel object. The columns show iCub executing the *push right* behavior which it successfully chose among the behaviors in its repertoire leading to the effect category "`moved-right`". **Second row:** Similarly, iCub interprets the interaction in Figures 9(e)-9(f)) as an instance of "`moved left`" verb concept (i.e., $*^6 0^6\ *^3\ 0\ *^4\ 0[**0]^2 0^3\ *^2\ 0^4 * 0^{14}\ *^5\ 0^6 * 0\ *^2\ [0*]^2\ *^2\ 0^4 * 0^2 * 0 - 0 * 0^4$), and a new object is put in front of it. It then chose to execute the *push left* behavior to produce the same effect.

more naturally without the designer being worried about how a certain verb is executed by the robot. For better evaluation of our proposal, we compared our effect prototypes with naive prototypes and exemplar-based conceptualization in goal emulation and multi-step planning tasks. Our evaluation showed that the regular-expression like nature of our conceptualization proposal combined with Mahalanobis distance performs better than the alternatives considered in the article.

Our prototypical representation of concepts is novel in that they represent the overall feature distribution in a category in a compact and efficient manner. This has several advantages: (i) Unimportant features can be discarded in similarity computation as also argued by (Verguts et al., 2004). (ii) Feature elements can be grouped and segmented together; other inter-feature relations and dependencies can be easily interpreted and recovered. (iii) Such a symbolic condensed representation is very suitable for goal specification. With these advantages at hand, we have demonstrated the advantages of our prototype-based concepts over exemplar-based and naive-prototype-based verb concepts.

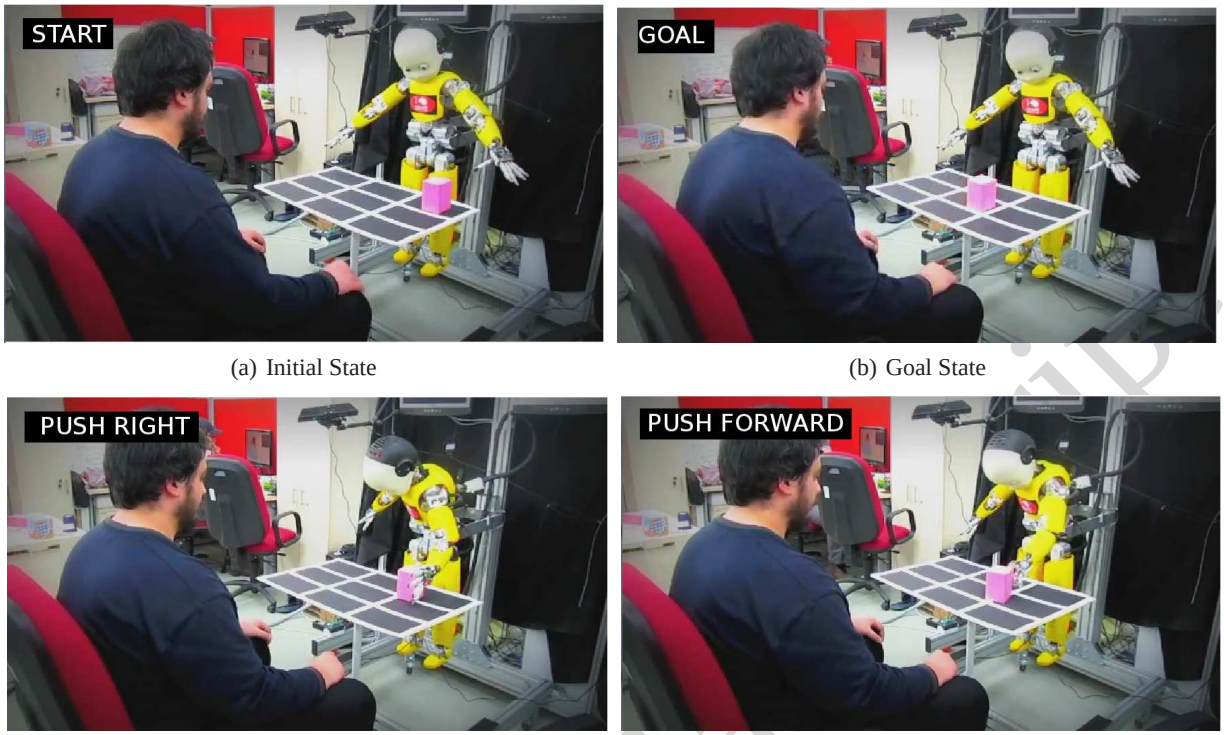
Verb concepts from effects: Biological Relevance

Our proposal of linking verb concepts to the effects of behaviors is in line with psychological ideomotor theories (e.g., Hommel et al. (2001)), according to which a behavior is represented in distal terms, *i.e.*, in terms of overall goals, not in proximal terms, *i.e.*, in terms of the kinematics of the movements and of the effectors required to reach the goal (see also (Hamilton et al., 2007)).

The neural underpinnings of this claim can be found in evidence on mirror neurons in monkeys, showing that they are activated preferentially when the behavior or the goal is clear (Umiltà et al., 2001, 2008). Such an association strongly indicates that the concept being conveyed by the verb is the request for a certain effect to be generated through the use of an appropriate behavior. In this sense, when we ask the robot to lift an object, we specify the goal as an increase of the object position in the vertical axis and leave the choice of the particular behavior to the robot itself. This is referred to as *goal emulation* in the literature as a form of imitation characterized by the replication of the observed end effect (Want & Harris, 2002), and is observed in infants after 12 months (Elsner, 2007).

Verb concepts from effects: Robotic and Computational Advantages

We claim that our proposal of linking verb concepts to the effects of behaviors and representing these concepts in terms of effect prototypes provides the following advantages, some



(a) Initial State (b) Goal State

(c) First Step (d) Second and Final Step

Figure 11. A sample execution of multi step planning using C_{EP} - *i.e.*, effect prototype based verb concepts. First, the initial state (a) and final state are shown (b). Then, the robot makes a plan involving *push-right* and *push-forward* behaviors, which are executed as shown in (c) and (d). When a simple Euclidian distance or naive prototype is used, the robot could not derive a plan.

of which have been demonstrated in this article:

- **Condensation:** The prototypes represent the distribution of features in a category using less storage. However, we have shown that, compared to exemplar-based conceptualization, this does not degrade the performance requirements expected from a humanoid (as listed below).
- **Low computational complexity:** A concept allows checking whether an item is of that concept or not. The fact that the information in a category is represented in a condensed manner facilitates faster checking of membership, hence faster interpretation of an observed event in terms of verb concepts.

The complexity of checking the membership of an effect instance f in n verb concepts is $O(n)$ in our proposal. However, that of exemplar-based conceptualization (C_{Ex}) is $O(n \times m)$, where m is the average number of items in a verb concept. The complexity of checking membership in the case of naive prototype conceptualization (C_{NP}) has also the complexity of $O(n)$; however, (i) the distance metric in Equation 9 requires more computations than the one in Equation 8 on average and (ii) Equation 9 leads to worse matching performance (as shown in Section "Results").

- **Flexibility:**

Our verb concepts provide flexibility in different aspects: (i) The same set of methods can be applied to another robot with a different embodiment having a different perceptual system and a different set of behaviors since the concepts are derived from the distribution of features and are not dependent on the set of behaviors and the features used. (ii) The prototypes allow a human to interact with the robot at a more symbolic and abstract level.

- **Robustness:**

Since in our proposal irrelevant changes in features are marked and not taken into consideration while interpreting effects and verb concepts, our prototype-based proposal of verb concepts is robust to changes in appearance and spatial changes, as demonstrated especially by the multi-step planning scenario where the alternatives failed to converge to a target state in 10 steps whereas our proposal (by comparing the most relevant and consistent features) converged to the target states in 2-3 steps.

The effect-prototype-based verb concepts, being an abstraction over behaviors, are beneficial for the following problems:

- **Goal specification and satisfaction:**

The robot provides flexibility to the user to provide commands at different levels: (i) at the language level, using verbs, in which case the robot can choose the best behavior that satisfies the corresponding effect, (ii) at the symbolic level, using strings of '+', '-', '0' and '*', making it easy for a human to specify in more detail what is expected to change, in which case the robot can again find the best behavior leading to the required change specification, (iii) at the low sensorimotor level, using exact values for features'

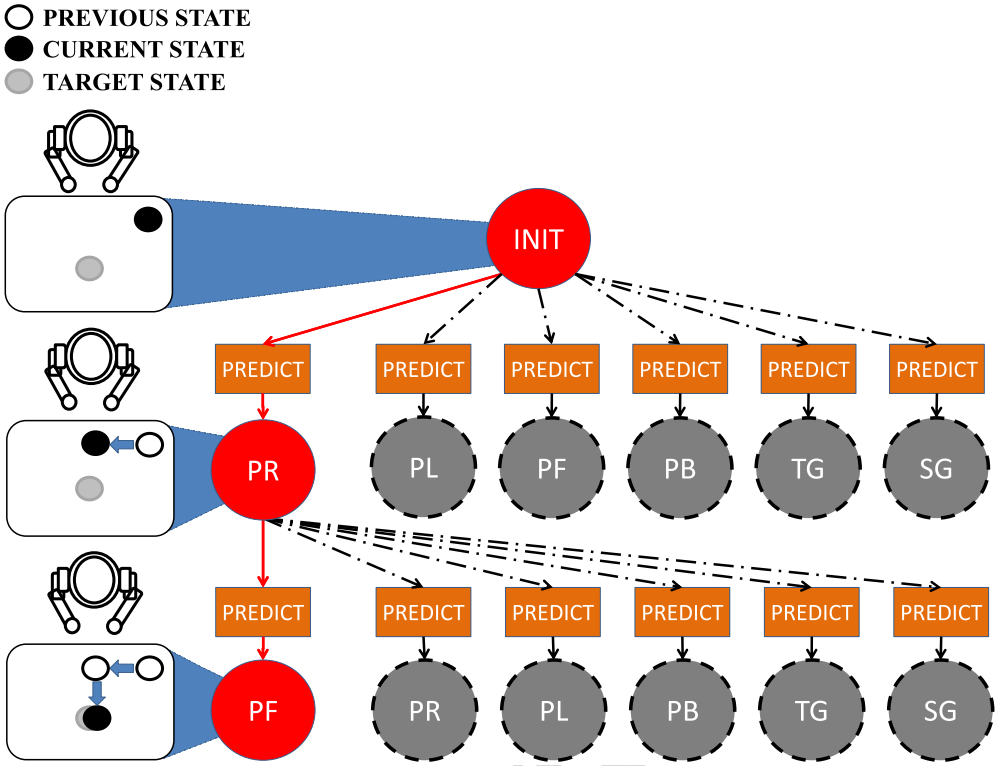


Figure 12. Multi-Step planning demonstration with effect prototype based verb concepts - C_{EP} (and the modified Mahalanobis distance in Equation 9). The behaviors are abbreviated as *PR* (*push-right*), *PL* (*push-left*), *PF* (*push-forward*), *PB* (*pull*), *TG* (*top-grasp*), *SG* (*side-grasp*). The planner successfully terminates with a reasonably small sequence of behaviors executed. The trial is also visualized in Figure 11.

final state, which again can be achieved using the conceptualization we have proposed. Moreover, as we have shown, the human can demonstrate an effect on any object and the robot can generate the same effect on a completely different object.

- **Language and human-robot interaction:**

An important cornerstone in language and seamless human-robot interaction is sharing the same meaning for the words that are used by humans and robots. With the verb concepts proposed in this article, we have addressed how verbs can be grounded in the sensorimotor system of the robot such that the robot can interpret in his own system the meaning associated with the word and utilize that meaning in various tasks involving interactions with humans.

Limitations and Future Directions

An important aspect of the system is the inclusion of supervision. The only supervision we put into the system is the effect labels that are provided by a human after each interaction. Although a developing infant gets such supervision throughout most of his development, it is worthwhile to investigate what the different effect categories could have been in the lack of supervision. The simplest idea would be to cluster the effect instances using an unsupervised clustering method. In (Akgün et al., 2009), we attempted an unsupervised approach to clustering the effect space; however, such an approach does not guarantee that the set of verb concepts would converge to be similar to the ones used by humans, and even if it did, it would take a longer time span. In a developing infant, both supervised and unsupervised mechanisms are used in the development of concepts, and we leave the integration of unsupervised categorization of effects as a future work. However, it should be noted that if we wish a robot to have the same concepts as we, humans, do, then we should provide supervision for the sensorimotor interactions.

For any computational system, representation is very important in that a suitable representation can simplify many tasks, and an unsuitable one can complicate many simple tasks unnecessarily. In fact, one can argue that cognitive development is about learning "suitable" representations from sensorimotor interactions. A representational requirement for our method is that the features extracted from the objects must be a fixed-length vector, and that the information extracted from the objects have fixed positions in this vector. For more complex objects, especially those with articulated parts, our methods can work with a hierarchical representation where the abstraction processes described in this article can be modified to work over the nodes of the hierarchy. For a scenario involving different behaviors or effects, a different set of features might be required to be able to represent the changes. However, the same abstraction process can be used as long as the features have fixed length and positions.

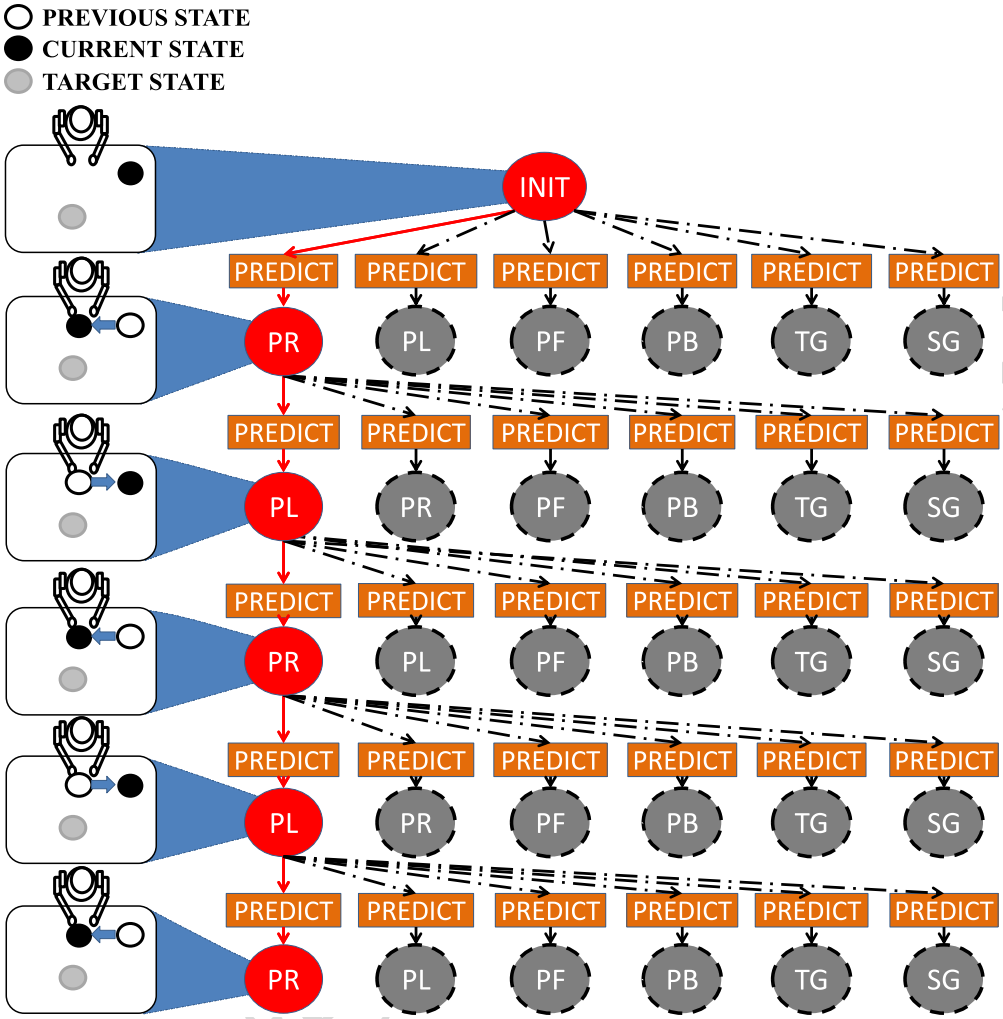


Figure 13. Multi-Step planning demonstration with naive prototype based verb concepts - C_{NP} (and the pure cMahalanobis distance in Equation 9). The behaviors are abbreviated as *PR* (*push-right*), *PL* (*push-left*), *PF* (*push-forward*), *PB* (*pull*), *TG* (*top-grasp*) and *SG* (*side-grasp*). The initial and the target states for the objects are the same with the one provided in Figure 11. Since the distance calculations yield wrong results due to irrelevant changes between the initial and goal states, the search does not terminate with success.

The current planning method is designed just to demonstrate the usefulness of verb concepts. Our planner makes plans in terms of "what" changes are required in the environment in order to reach to a target state, and finds a sequence of verb concepts to satisfy them. In a full-fledged cognitive system, the planning must be able to take into account also "how" some changes are performed in the environment. This can be achieved by having the behaviors parametric such that the same behavior with different parameters can yield different effects. The planner then can treat different parameter settings as different behaviors while making plans and determine the behavior with the parameters conforming to the required task.

Acknowledgments

This work is partially funded by the EU projects ROSSI (FP7-ICT-216125), and RobotCub (FP6-ICT-004370), and by TÜBİTAK (Turkish Scientific and Technology Council) through projects no 109E033 and 111E287.

Footnotes

[1]In this paper, we will assume that the subject is given through gaze or other means.

[2]For simplicity, assume that the object is pointed through mechanisms such as shared gaze.

[3]The reaching part of these behaviors is achieved using a modified form of Dynamic Movement Primitives (Akgün et al., 2010) and the remaining parts of the behaviors are pre-coded. Due to Dynamic Movement Primitives, there is a feedback loop in the system allowing the robot to adapt to changes in position.

References

Akgün, B., Dağ, N., Bilal, T., Atıl, I., & Şahin, E. (2009). Unsupervised learning of affordance relations on a humanoid robot. *24th*

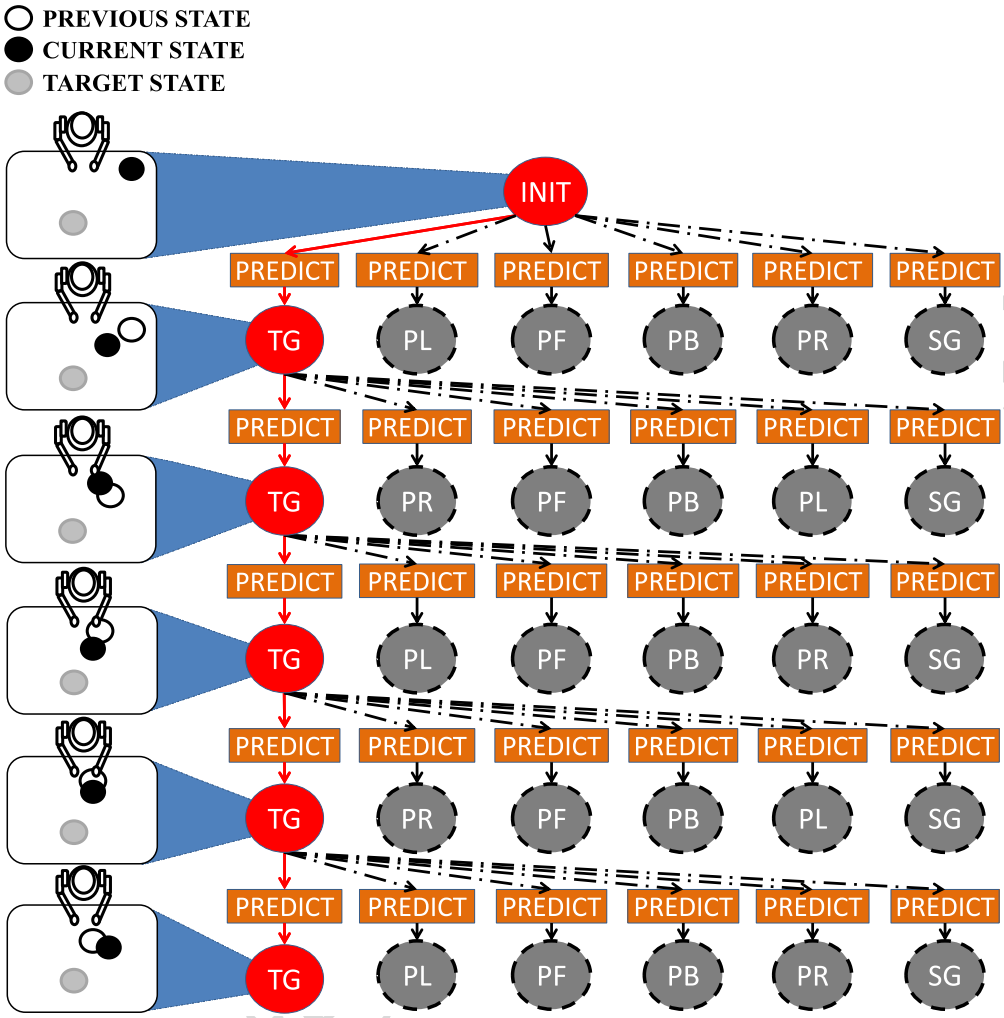


Figure 14. Multi-Step planning demonstration with examplar-based verb concepts - C_{Ex} (and Euclidean distance in Equation 10). The behaviors are abbreviated as *PR* (*push-right*), *PL* (*push-left*), *PF* (*push-forward*), *PB* (*pull*), *TG* (*top-grasp*) and *SG* (*side-grasp*). The initial and the target states for the objects are the same with the one provided in Figure 11. Since the distance calculations yield wrong results due to irrelevant changes between the initial and goal states, the search does not terminate with success.

International Symposium on Computer and Information Sciences (ISCIS), 254–259.

Akgün, B., Tunaoglu, D., & Sahin, E. (2010). Action recognition through an action generation mechanism. *International Conference on Epigenetic Robotics*.

Alissandrakis, A., Nehaniv, C., & Dautenhahn, K. (2003). Synchrony and perception in robotic imitation across embodiments. In *Computational intelligence in robotics and automation, 2003. proceedings. 2003 ieee international symposium on* (Vol. 2, pp. 923–930).

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, *37*(3), 372–400.

Barsalou, L. (1999). Perceptual symbol systems. *Behavioral and brain sciences*, *22*(4), 577–660.

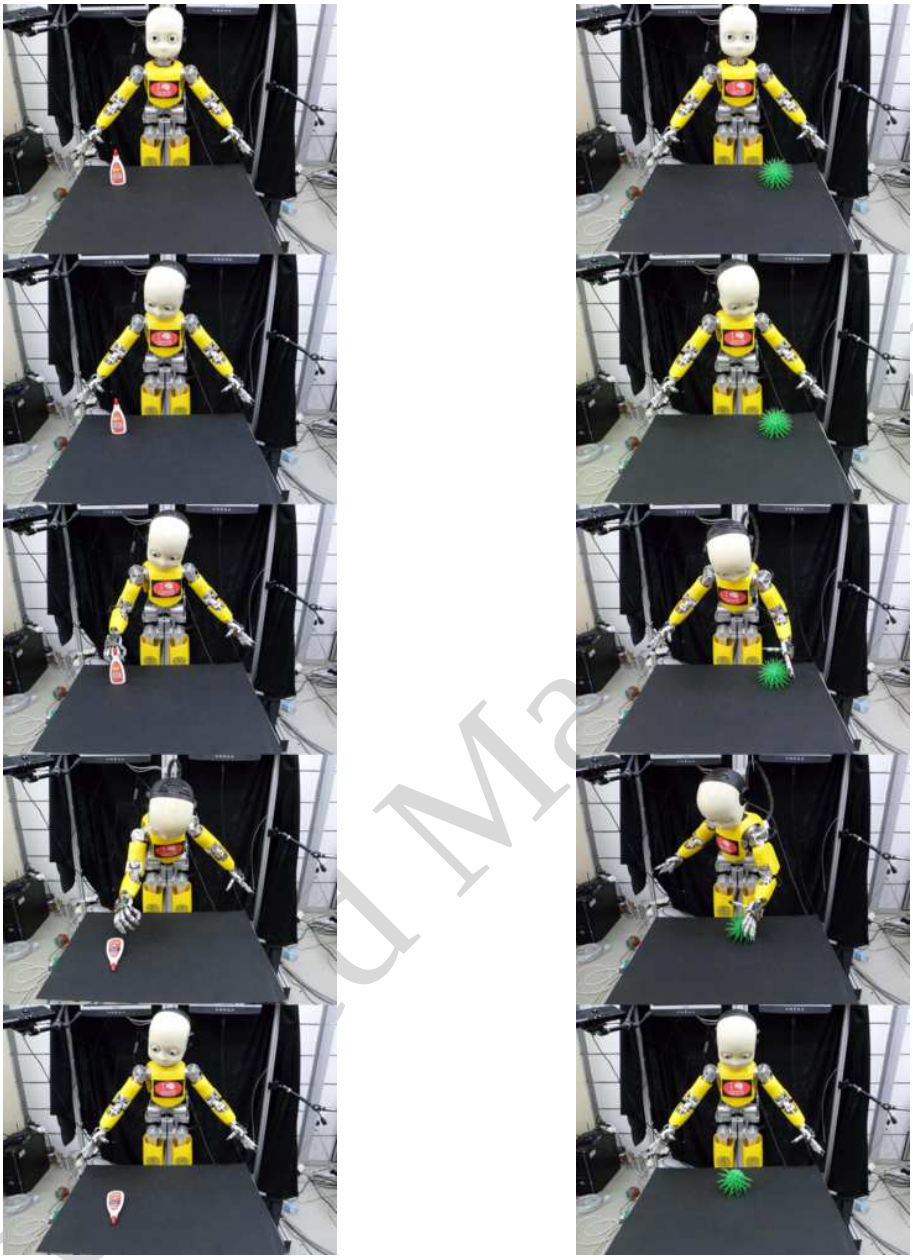
Bergquist, T., Schenck, C., Ohiri, U., Sinapov, J., Griffith, S., & Stoytchev, A. (2009). Interactive object recognition using proprioceptive feedback. *IROS Workshop: Semantic Perception for Mobile Manipulation*.

Borghi, A. M. (2007). Object concepts and embodiment: Why sensorimotor and cognitive processes cannot be separated. *La nuova critica*, *15*(4), 447-472.

Borghi, A. M. (2012). Action language comprehension, affordances and goals. In Y. Coello & A. Bartolo (Eds.), *Language and action in cognitive neuroscience. contemporary topics in cognitive neuroscience series* (pp. 125–143). Psychology Press.

Borghi, A. M., & Riggio, L. (2009). Sentence comprehension and simulation of object temporary, canonical and stable affordances. *Brain Research*, *1253*, 117–128.

Bruner, J., Goodnow, J., & Austin, G. (1986). *A study of thinking*. Transaction Publishers.



(a) $*^{26}0^3*[**000]^20^{12}*^40*^5[*0]^2*^6 0^5*^3-0^2*0^4$

(b) $*^{16}0*0*^{18}0^{15}*^{13}0*^60*0*0*^4 + *^5 0$

Figure 15. Goal specification with symbols. **(a)** iCub is given a goal $*^{80} - *^60$ (For the sake of space, we denote k consecutive occurrences of a symbol s with s^k), meaning that it should produce a decrease in the x position while avoiding a change in object presence (*i.e.*, the object should not disappear) and the change in the other dimensions can be ignored. iCub matches this goal with the *moved forward* verb concept and chooses to execute the *push forward* behavior accomplishing the specified goal. **(b)** Similarly, iCub is given a goal $*^{81} - *^50$, meaning that iCub is to produce an increase in the y position of the object and the change in the other dimensions can be ignored while, again, the object presence should not change. iCub matches this goal with the *moved right* verb concept and executes *push right* behavior to accomplish the goal.

Cangelosi, A. (2001). Evolution of communication and language using signals, symbols, and words. *IEEE Transactions on Evolutionary Computation*, *5*(2), 93–101.

Cangelosi, A. (2010). Grounding language in action and perception: From cognitive agents to humanoid robots. *Physics of Life Reviews*, *7*(2), 139-151.

Cangelosi, A., & Harnad, S. (2001). The adaptive advantage of symbolic theft over sensorimotor toil: Grounding language in

perceptual categories. *Evolution of Communication*, *4*(1), 117–142.

Cangelosi, A., Hourdakis, E., & Tikhanoff, V. (2006). Language acquisition and symbol grounding transfer with neural networks and cognitive robots. *International Joint Conference on Neural Networks (IJCNN)*, 1576–1582.

Cangelosi, A., Metta, G., Sagerer, G., Nolfi, S., Nehaniv, C., Fischer, K., et al. (2010). Integration of Action and Language Knowledge: A Roadmap for Developmental Robotics. *IEEE Transactions on Autonomous Mental Development*, *2*(3), 167–195.

Cangelosi, A., & Parisi, D. (2004). The processing of verbs and nouns in neural networks: Insights from synthetic brain imaging. *Brain and Language*, *89*(2), 401–408.

Cangelosi, A., & Riga, T. (2006). An embodied model for sensorimotor grounding and grounding transfer: Experiments with epigenetic robots. *Cognitive science*, *30*(4), 673–689.

Christiansen, M., & Kirby, S. (2003). Language evolution: Consensus and controversies. *Trends in Cognitive Sciences*, *7*(7), 300–307.

Cohen, P., Morrison, C., & Cannon, E. (2005). Maps for verbs: The relation between interaction dynamics and verb use. In *Proceedings of the 9th international conference on artificial intelligence (ijcai)*.

Elsner, B. (2007). Infants' imitation of goal-directed actions: the role of movements and action effects. *Acta psychologica*, *124*(1), 44–59.

Fischer, M., & Zwaan, R. (2008). Embodied language: A review of the role of the motor system in language comprehension. *The Quarterly Journal of Experimental Psychology*, *61*(6), 825–850.

Gabora, L., Rosch, E., & Aerts, D. (2008). Toward an ecological theory of concepts. *Ecological Psychology*, *20*(1), 84–116.

Gallese, V., & Lakoff, G. (2005). The brain's concepts: The role of the sensory-motor system in conceptual knowledge. *Cognitive neuropsychology*, *22*(3), 455–479.

Gärdenfors, P. (2004). *Conceptual spaces: The geometry of thought*. The MIT Press.

Gibson, J. J. (1986). *The ecologial approach to visual perception*. Lawrence Erlbaum Associates.

Glenberg, A., & Kaschak, M. (2002). Grounding language in action. *Psychonomic Bulletin & Review*, *9*(3), 558.

Glenberg, A., & Robertson, D. (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language*, *43*(3), 379–401.

Glenberg, A., Sato, M., Cattaneo, L., Riggio, L., Palumbo, D., & Buccino, G. (2008). Processing abstract language modulates motor system activity. *The Quarterly Journal of Experimental Psychology*, *61*(6), 905–919.

Hamilton, A., Grafton, S., & Hamilton, A. (2007). The motor hierarchy: from kinematics to goals and intentions. In P. Haggard, Y. Rossetti, & M. Kawato (Eds.), *Sensorimotor foundations of higher cognition, attention and performance* (pp. 381–408). Oxford University Press.

Harnad, S. (1990). The symbol grounding problem. *Physica*, *D*(42), 335-346.

Hashimoto, T., & Masumi, A. (2007). Learning and transition of symbols: Towards a dynamical model of a symbolic individual. In C. N. C. Lyon & A. Cangelos (Eds.), *Emergence of communication and language* (pp. 223–236). Springer.

Hommel, B., Müsseler, J., Aschersleben, G., & Prinz, W. (2001). The theory of event coding (tec): A framework for perception and action planning. *Behavioral and Brain Sciences*, *24*(05), 849–878.

Jebara, T. (2004). *Machine learning: discriminative and generative* (Vol. 755). Springer.

Johansen, M., & Kruschke, J. (2005). Category representation for classification and feature inference. *Learning, Memory*, *31*(6), 1433–1458.

Kozima, H., Nakagawa, C., & Yano, H. (2002). Emergence of imitation mediated by objects. *Lund University Cognitive Studies*, 59-61.

Krunic, V., Salvi, G., Bernardino, A., Montesano, L., & Santos-Victor, J. (2009). Affordance based word-to-meaning association. *IEEE Int. Conference on Robotics and Automation (ICRA)*, 4138–4143.

Kruschke, J. (2005). Category learning. *The handbook of cognition, Eds: K. Lamberts, R. L. Goldstone*, 183–201.

Leopold, D., O'Toole, A., Vetter, T., & Blanz, V. (2001). Prototype-referenced shape encoding revealed by high-level aftereffects. *Nature Neuroscience*, *4*(1), 89–94.

Lyon, C., Nehaniv, C., & Cangelosi, A. (2007). *Emergence of communication and language*. Springer-Verlag New York Inc.

Mahalanobis, P. (1936). On the generalized distance in statistics. In *Proceedings of the national institute of sciences of india* (Vol. 2, pp. 49–55).

Marocco, D., Cangelosi, A., Fischer, K., & Belpaeme, T. (2010). Grounding action words in the sensorimotor interaction with the world: experiments with a simulated icub humanoid robot. *Frontiers in Neurorobotics*, *4*(7), 1-15.

Metta, G., & Fitzpatrick, P. (2003). Better vision through manipulation. *Adaptive Behavior*, *11*(2), 109-128.

Metta, G., Sandini, G., Vernon, D., Natale, L., & Nori, F. (2008). The iCub humanoid robot: an open platform for research in embodied cognition. In *Proceedings of the 8th workshop on performance metrics for intelligent systems* (pp. 50–56).

Minda, J., & Smith, J. (2001). Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(3), 775–799.

Montesano, L., Lopes, M., Bernardino, A., & Santos-Victor, J. (2008). Learning object affordances: From sensory–motor coordination to imitation. *IEEE Transactions on Robotics*, *24*(1), 15–26.

Montesano, L., Lopes, M., Melo, F., Bernardino, A., & Santos-Victor, J. (2009). A computational model of object affordances. *Advances in Cognitive Systems*.

Nehaniv, C. L., Lyon, C., & Cangelosi, A. (2007). Current work and open problems: A road-map for research into the emergence of communication and language. In C. L. N. C. Lyon & A. Cangelosi (Eds.), *Emergence of communication and language* (pp. 1–27). Springer.

Nosofsky, R., Kruschke, J., & McKinley, S. (1992). Combining exemplar-based category representations and connectionist learning rules. *Learning, Memory*, *18*(2), 211–233.

Nosofsky, R., & Zaki, S. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Learning, Memory*, *28*(5), 924–940.

Parthemore, J., & Morse, A. (2010). Representations reclaimed: Accounting for the co-emergence of concepts and experience. *Pragmatics & Cognition*, *18*(2), 273–312.

Qin, A., & Suganthan, P. (2004). Robust growing neural gas algorithm with application in cluster analysis. *Neural Networks*, *17*(8-9), 1135–1148.

Rosch, E. (1973). Natural categories. *Cognitive psychology*, *4*(3), 328–350.

Rosseel, Y. (2002). Mixture Models of Categorization. *Journal of Mathematical Psychology*, *46*(2), 178–210.

Rouder, J., & Ratcliff, R. (2006). Comparing exemplar-and rule-based theories of categorization. *Current Directions in Psychological Science*, *15*(1), 9-13.

Rudolph, M., Muhlig, M., Gienger, M., & Bohme, H.-J. (2010). Learning the consequences of actions: Representing effects as feature changes. *Int. Symposium on Learning and Adaptive Behavior in Robotic Systems*.

Rusu, R. B., & Cousins, S. (2011). 3d is here: Point cloud library (pcl). *Library*, *26*(2), 1-4.

Sahin, E., Çakmak, M., Doğar, M., Uğur, E., & Üçoluk, G. (2007). To afford or not to afford: A new formalization of affordances toward affordance-based robot control. *Adaptive Behavior*, *15*(4), 447–472.

Steels, L. (2003). Evolving grounded communication for robots. *Trends in Cognitive Science*, *7*(7), 308–312.

Steels, L. (2007). The recruitment theory of language origins. In C. L. N. C. Lyon & A. Cangelosi (Eds.), *Emergence of communication and language* (pp. 129–150). Springer.

Uğur, E., Şahin, E., & Öztop, E. (2009). Affordance learning from range data for multi-step planning. *9th International Conference on Epigenetic Robotics (Epirob)*, *146*, 177–184.

Uğur, E., & Şahin, E. (2010). Traversability: A case study for learning and perceiving affordances in robots. *Adaptive Behavior*, *18*(3-4), 258–284.

Umiltà, M., Intskirveli, I., Grammont, F., Rochat, M., Caruana, F., Jezzini, A., et al. (2008). When pliers become fingers in the monkey motor system. *Proceedings of the National Academy of Sciences*, *105*(6), 2209.

Umiltà, M., Kohler, E., Gallese, V., Fogassi, L., Fadiga, L., Keysers, C., et al. (2001). I know what you are doing: A neurophysiological study. *Neuron*, *31*(1), 155–165.

Verguts, T., Ameel, E., & Storms, G. (2004). Measures of similarity in models of categorization. *Memory & Cognition*, *32*(3), 379.

Want, S. C., & Harris, P. L. (2002). How do children ape? Applying concepts from the study of non-human primates to the developmental study of imitation in children. *Developmental Science*, *5*(1), 1–13.

Zwaan, R., & Taylor, L. (2006). Seeing, acting, understanding: Motor resonance in language comprehension. *Journal of Experimental Psychology-General*, *135*(1), 1–11.