Beyond the Self: Using Grounded Affordances to Interpret and Describe Others' Actions

Giovanni Saponaro, *Student Member, IEEE*, Lorenzo Jamone, *Member, IEEE*, Alexandre Bernardino, *Senior Member, IEEE*, and Giampiero Salvi

***Abstract*—In this paper, we propose a developmental approach that allows a robot to interpret and describe the actions of human agents by reusing previous experience. The robot first learns the association between words and object affordances by manipulating the objects in its environment. It then uses this information to learn a mapping between its own actions and those performed by a human in a shared environment. It finally fuses the information from these two models to interpret and describe human actions in light of its own experience. In our experiments, we show that the model can be used flexibly to do inference on different aspects of the scene. We can predict the effects of an action on the basis of object properties. We can revise the belief that a certain action occurred, given the observed effects of the human action. In an early action recognition fashion, we can anticipate the effects when the action has only been partially observed. By estimating the probability of words given the evidence and feeding them into a predefined grammar, we can generate relevant descriptions of the scene. We believe that this is a step toward providing robots with the fundamental skills to engage in social collaboration with humans.**

***Index Terms*—Affordances, embodied cognition, gestures, humanoid robots, language acquisition through development.**

I. Introduction

COOPERATION, or the ability of working successfully in groups, is a tenet of human society [1]. This skill is acquired by human children incrementally, around the second year of life, as they develop the ability to coordinate themselves with peers or adult caregivers in shared problem-solving activities and social games [2]. This is achieved not only by mere behavioral coordination, but also by employing communicative strategies [3] and by continuously observing partners' actions [4]. Loosely inspired by these observations, this paper presents and evaluates a cognitive system for robots which permits reasoning over subsequent phases: first about self-learned knowledge (about affordances and language-based descriptions of objects), and then about others' actions.

Even though social robots[1] are becoming common in domestic and public environments, human–robot teams still lag behind human–human teams in terms of effectiveness. For robots, interpreting the actions of others and learning to describe them verbally (for effective cooperation) is challenging. The reason is that we cannot possibly model all the imaginable physical, verbal, and nonverbal (e.g., gestures) cues that can take place during human–robot interaction, due to the richness of language and the high variability of the real world outside of structured research laboratories and factories. Hence, it is necessary to have robots that *learn* world elements and properties of language [6], and the ability to link these verbal elements with other skills, such as other perceptual modalities (e.g., vision of objects and other agents) and manipulation abilities (e.g., grasping objects and placing them in order to achieve a goal) [7].

This paper builds upon the intuition that a robot can generalize its previously acquired knowledge of the world (e.g., motor actions, objects properties, physical effects, and verbal descriptions) to those situations where it observes a human agent performing familiar actions in a shared human–robot scenario. We follow the developmental robotics perspective [8], [9], which takes inspiration from the progressive learning phenomena observed in children's mental development (e.g., the understanding of language, the acquisition of manipulation skills, and the comprehension of others' actions), and investigates how to model the evolution and acquisition of these increasingly complex cognitive processes in artificial autonomous systems.

In particular, we are inspired by the possible existence of a shared representation for self-related and others-related knowledge in the human brain [10]–[12], and we look at the developmental stages in which human children have consolidated an idea of self–other distinction [13] and start to reason about the external world also in allocentric terms [14], in

Manuscript received November 15, 2017; revised September 6, 2018; accepted November 14, 2018. Date of publication January 17, 2019; date of current version June 10, 2020. This work was supported in part by Fundação para a Ciência e a Tecnologia Projects under Grant UID/EEA/50009/2013 and Grant AHA CMUP-ERI/HCI/0046/2013, and in part by CHIST-ERA Project IGLU. *(Corresponding author: Giovanni Saponaro.)*

G. Saponaro and A. Bernardino are with the Institute for Systems and Robotics, Instituto Superior Técnico, Universidade de Lisboa, 1049-001 Lisbon, Portugal (e-mail: gsaponaro@isr.tecnico.ulisboa.pt; alex@isr.tecnico.ulisboa.pt).

L. Jamone is with the Advanced Robotics at Queen Mary, School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, U.K., and also with the Institute for Systems and Robotics, Instituto Superior Técnico, Universidade de Lisboa, 1049-001 Lisbon, Portugal (e-mail: l.jamone@qmul.ac.uk).

G. Salvi is with the School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, 10044 Stockholm, Sweden (e-mail: giampi@kth.se).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCDS.2018.2882140

[1] A social robot is "[a robot that is] able to communicate and interact with us, understand and even relate to us, in a personal way. [It] should be able to understand us and itself in social terms" [5].

addition to the ego-centric ones, and could therefore possibly begin to use knowledge about the *self* to infer about *others*.

Extending on our recent work [15], in this paper we combine robot ego-centric learning about language and object affordances [16] with the observation of external agents through gesture recognition [17]. Our novel contributions are as follows.

1) A probabilistic method to fuse self-learned knowledge of language and object affordances, with socially aware information of others' physical actions (in the form of uncertain soft evidence).
2) Experimental findings showing the reasoning power of our combined system, which is able to make inferences and predictions over affordances and words.
3) The possibility of generating verbal descriptions from the estimated word probabilities and a predefined grammar, with emergence of nontrivial language properties such as congruent/incongruent conjunctions, synonyms between two consecutive sentences speaking about the same concepts.

Furthermore, we make our human action data and probabilistic reasoning code publicly available[2,3] in the interest of reproducibility.

This paper is structured as follows. In Section II, we briefly overview the literature on the interpretation and verbal description of others in different disciplines. In Section III, we present our proposed method and its components. In Section IV, we provide details and assumptions of the approach. Section V illustrates our results, and in Section VI, we draw our concluding remarks.

II. Related Work

Human cooperation is a phenomenon that we often take for granted (at least in adults), possibly because it is widespread and intimately embedded into human societies. However, this nontrivial skill is greatly facilitated, and influenced, by human language [18]. For instance, educational research has shown that, when language is used as a cultural tool for intellectual tasks in preteen students, discursive interaction enables collective thinking to become more effective, also fostering individual reasoning and faster learning [19].

The ability to understand and interpret our peers has also been studied in neuroscience and psychology, focusing on internal simulations and re-enactments of previous experiences [20], [21], or on visuomotor neurons [11], i.e., neurons that are activated by visual stimuli. Mirror neurons respond to action and object interaction, both when the agent acts and when it observes the same action performed by others, hence the name "mirror." They are based on the principle that perceptual input can be linked with the human action system for predicting future outcomes of actions, i.e., the effect of actions, particularly when the person possesses concrete prior personal experience of the actions being observed in others [22], [23].

In applying the mirror neuron theory in robotics, as we and others do [24], [25], an agent can first acquire knowledge by sensing and self-exploring its surrounding environment. Afterwards, it can employ that learned knowledge to novel observations of another agent (e.g., a human person) who performs similar physical actions to the ones executed during prior training. In particular, when the two interacting agents are a caregiver and an infant, the mechanism is called *parental scaffolding*, having been implemented on robots too [26], [27]. These works tackle the so-called correspondence problem [28], in our case in a simple collaboration scenario, assuming that the two agents are capable of applying actions to objects leading to similar effects, enabling the transfer, and that they operate on a shared space (i.e., a table accessible by both agents' arms). The morphology and the motor realization of the actions can be different between the two agents.

Some authors have studied the ability to interpret other agents under the deep learning paradigm. In [29], a recurrent neural network is proposed to have an artificial simulated agent infer human intention (as output) from joint input information about objects, their potential affordances or opportunities, and human actions, employing different time scales for different actions. However, in that work a virtual simulation able to produce large quantities of data was used. This is both unrealistic when trying to explain human cognition, and limited, because a simulator cannot model all the physical events and the unpredictability of the real world. In contrast, we use real, noisy data acquired from robots and sensors to validate our model. In addition, deep neural networks trained with large amounts of data can be difficult to inspect in their inner layers and activations [30], whereas our Bayesian model is focused on exhibiting emerging patterns of causality, choices, explanations from relatively few data points.

DeepMind and Google published a method [31] to perform relational reasoning on images, i.e., a system that learns to reflect about entities and their mutual relations, with the ability of providing answers to questions such as "Are there any rubber things that have the same size as the yellow metallic cylinder?" That work is very powerful from the point of view of cognitive systems, vision, and language. Our approach is different because: 1) we focus on *robotic* cognitive systems, including manipulation and the uncertainties inherent to robot vision and control and 2) we follow the developmental paradigm and the embodiment hypothesis [8], meaning that, leveraging the fact that a human and a humanoid produce actions with similar effects, we relate words with the robot's *sensorimotor* experience, rather than sensory only (purely images-to-text).

In robotics and cognitive systems research, both object-directed action recognition in external agents [32] and the incorporation of language in human–robot systems [33], [34] have received ample attention, for example using the concept of *intuitive physics* [35], [36] to be able to predict outcomes from real or simulated interactions with objects. A growing interest is devoted to robots that learn new cognitive skills and improve their capabilities by interacting autonomously with the surrounding environment. Robots operating in the real, unstructured world may understand available opportunities conditioned on their body, perception, and sensorimotor experiences: the intersection of these elements gives rise to

[2]https://github.com/giampierosalvi/AffordancesAndSpeech: the code from [16] has been extended to support the experiments in this paper.

[3]https://github.com/gsaponaro/tcds-gestures: code from this paper.

object *affordances* (action possibilities), as they are called in psychology [37]. The advantage of robot affordances lies in the ability to capture essential functional properties of environment objects in terms of the actions that the agent is able to perform with them, allowing to reason with prior knowledge about never-before-seen scenarios, thus exhibiting learning [38], [39] and some degree of online adaptation [40].

Zech *et al.* [41] published a systematic taxonomy of robot affordance models. According to their criteria (we refer the reader to the taxonomy for the precise definitions), in terms of *perception* this paper classifies as using an agent perspective, meso-level features, first order, and stable temporality; in terms of *development*: acquisition by exploration, prediction by inference, generalization exploitation by action selection and language, and offline learning.

Several works have studied the potential coupling between learning robot affordances and *language grounding*. The union of these two elements can give new skills to cognitive robots, such as: creation of categorical concepts from multimodal association obtained by grasping and observing objects, while listening to partial verbal descriptions [42], [43]; associating spoken words with sensorimotor experience [16], [44]; linking language with sensorimotor representations [45]; or carrying out complex tasks (which require planning of a sequence of actions) expressed in natural language instructions to a robot [46].

In particular Salvi *et al.* [16], which this paper extends, proposed a joint model to learn robot affordances (i.e., relationships between actions, objects, and resulting effects) together with word meanings. The data used for learning such a model is from robot manipulation experiments, acquired from an ego-centric perspective. Each experiment is associated with a number of alternative verbal descriptions uttered by two human speakers, for a total of 1270 recordings. That framework assumes that the robot action is known *a priori* during the training phase (e.g., during a grasping action the robot knows with certainty that it is performing a grasp), and the resulting model can be used at testing to make inferences about the environment. In a recent work [15], we relaxed the assumption of knowing the action. We did this by merging the action estimation obtained from an external gesture recognizer [17] as *hard evidence* (i.e., certain evidence) to the full model, meaning that the action was deterministic. By contrast, in this paper we propose a theoretical way to fuse the two sources of information (about the self and about others) in a fully probabilistic manner, therefore introducing *soft evidence*. This addition allows to perform more fine-grained types of inferences and reasoning than before. First, predictions over affordances and words when observing another agent with uncertainty. Second, the generation of *verbal descriptions* from the estimated word probabilities, for easier human interpretation of the model's explanations.

III. Method

The purpose of this paper is to model the development of language learning from self-centered, individualistic learning to socially aware learning. This transition happens gradually in subsequent phases. In the first phase, the system engages in manipulation activities with objects in its environment [38]. The robot learns object affordances by associating object properties, actions, and the corresponding effects. In a second phase, the robot interacts with a human who uses spoken language to describe the robot's activities [16]. Here, the robot interprets the meaning of the words, grounding them in the action–perception experience acquired so far. Although this phase can already be considered *social* for the presence of a human *narrator*, it is still self-centered, because the robot is still learning how to interpret its own actions. In the last phase, which is the contribution of this paper, the system turns to observing human actions of a similar nature as the ones explored in the first phases (see Fig. 1). The robot reuses the experience acquired in the first phases to interpret the new observations and to address the correspondence problem [28] between its own actions and the actions performed by the human. In this phase, human movements are interpreted using the experience acquired so far, and they are incorporated into the model using a statistical gesture recognizer [17].

Fig. 2 illustrates the probabilistic dependencies in the complete model and will be detailed in the following sections.

To permit the transfer from robot self-centered knowledge to human knowledge to work, we assume that the *same actions*, performed on objects with the *same properties*, cause the *same effects* and are described by the *same words*. In other terms, all of the variables under consideration (which will be described in Section IV) are the link between robot and human.

In our theoretical formulation and in our implementation, we will hinge on the existence of the discrete Action variable, the value of which is known to the robot in the ego-centric phase of learning, but must be inferred when observing human actions. This variable connects all the other observable variables in the model: human gesture features, object properties, effect variables, and words. This allows the robot to:

1) use language in order to determine the mapping between human and own actions, and learn the corresponding perceptual models;
2) in many cases, use the affordance variables to infer the above mapping even in the absence of verbal descriptions;
3) once the perceptual models for human actions are acquired, use the complete model to do inference on any variable given some evidence.

In the remainder of this section, first we provide details, in Section III-A, about the probabilistic models enclosed in the *affordance–words model* box of Fig. 2. Then, in Section III-B, we describe the gesture recognition method. Finally, in Section III-C, we describe the way in which we combine evidence from the two models.

A. Affordance–Words Model

We use a Bayesian probabilistic framework to allow a robot to ground the basic world behavior and verbal descriptions associated to it. All variables in the model are discrete or are discretized from continuous sensory variables through clustering in a preliminary learning phase. The variables can be divided according to their use: action variable $A = \{a\}$,

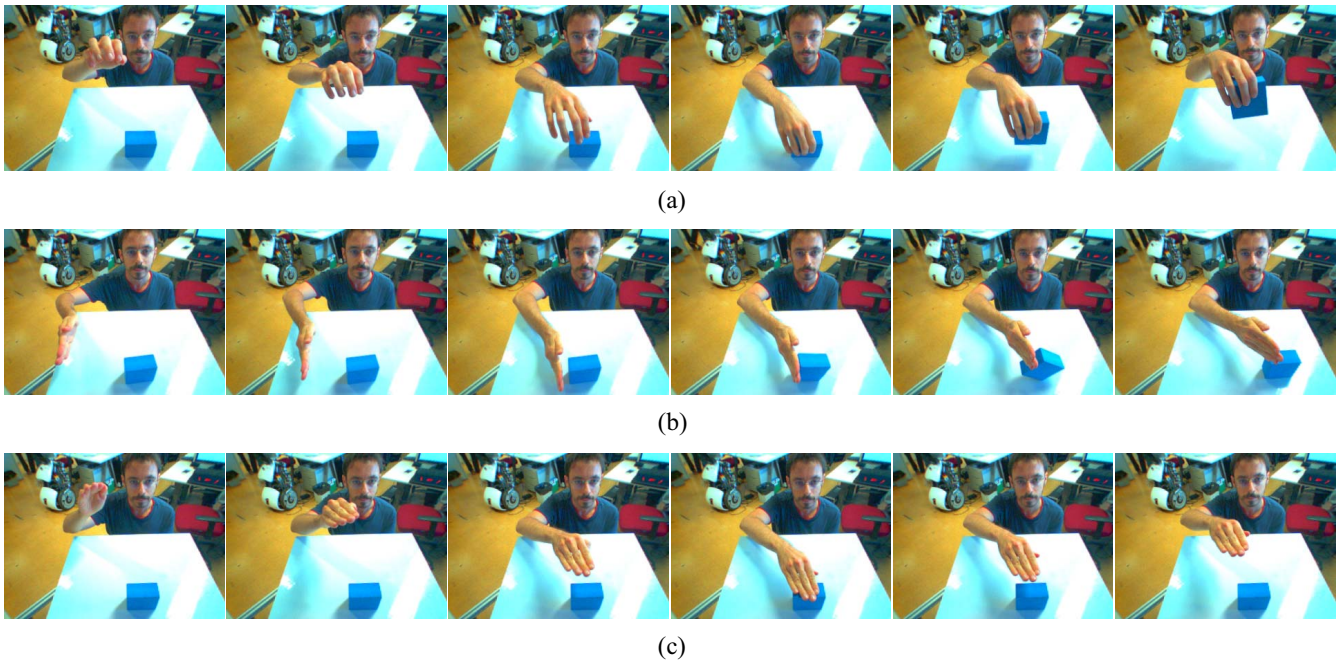


Fig. 1. Examples of human actions from the point of view of the robot. (a) Grasp action: moving the hand toward an object vertically, then grasping and lifting it. (b) Tap action: moving the hand toward an object laterally then touching it, causing a motion effect. (c) Touch action: moving the hand toward an object vertically, touching it (without grasping), then retracting the hand.

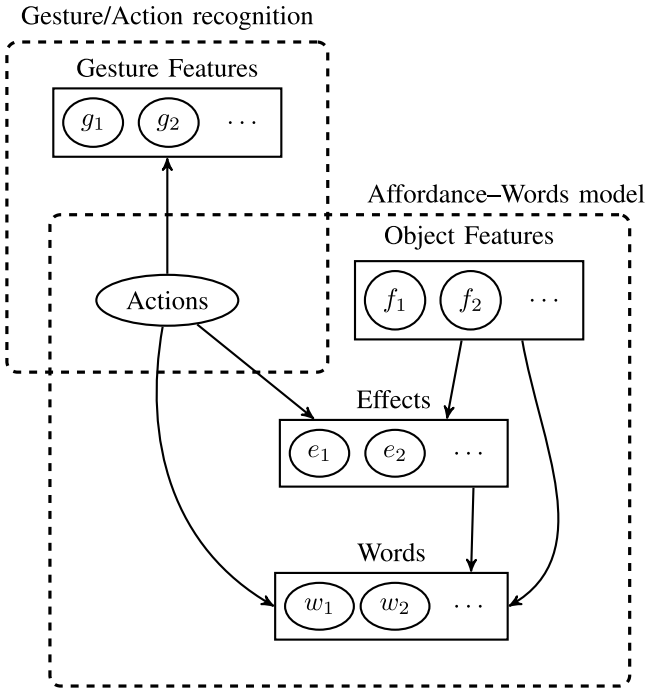


Fig. 2. Abstract representation of the probabilistic dependencies in the model.

object feature variables $F = \{f_1, f_2, \ldots\}$, effect variables $E = \{e_1, e_2, \ldots\}$, and word variables $W = \{w_1, w_2, \ldots\}$. Details on the specific variables used in this paper are given in Section IV.

The Bayesian network (BN) model [47] relates all these variables by means of the joint probability distribution $P_{\text{BN}}(A, F, E, W)$, sketched by the affordance–words model box in Fig. 2. The dependency structure and the model parameters are estimated by the robot in an ego-centric way through interaction with the environment. As a consequence, during learning, the robot knows what action it is performing with certainty, and the variable A assumes a deterministic value. During inference, the probability distribution of the variable A can be inferred from evidence on the other variables. For example, if the robot is asked to make a spherical object roll, it will be able to select the action tap as most likely to obtain the desired effect, based on previous experience.

B. Gesture Recognition

When observing a human performing an action, the value of the variable A is not known to the robot neither during learning nor during inference. During learning, we assume that the robot has not yet acquired a perceptual model of the gestures associated to the human actions. However, the value of A can be inferred, either from a verbal description of the scene, or from the other affordance variables through the affordance–words model described earlier.

For example, suppose that the affordance–words model predicts that performing a tap action on a spherical object will result in a high velocity of the object. If the human performs an unknown action on a spherical object and obtains a high velocity, the robot will be able to infer that the action is most probably a tap, although it was not able to recognize the gesture associated with this action.

This information can be used to train our statistical *gesture recognition system* [17]. The system recognizes actions (from gesture features) and corresponds to the gesture/action recognition block in Fig. 2. It is based on Hidden markov models

(HMMs) with Gaussian mixture models as emission probability distributions. Our input features are the 3-D coordinates of the tracked human hand indicated by the g_i variables in Fig. 2. The coordinates are transformed to be centered on the person torso (to be invariant to the distance between the user and the sensor) and normalized to account for variability in amplitude (to be invariant to wide/emphatic vs narrow/subtle executions of the same action).

The model for each action is a left-to-right HMM, where the transition model between the Q discrete states $\mathcal{S} = \{s_1, \ldots, s_Q\}$ is structured so that states with a lower index represent events that occur earlier in time.

Although not expressed so far in the notation, the continuous variables g_i are measured at regular time intervals. At a certain time step t, the D-dimensional feature vector can be expressed as $\boldsymbol{g}[t] = \{g_1[t], \ldots, g_D[t]\}$. The input to the model is a sequence of T such feature vectors $\boldsymbol{g}[1], \ldots, \boldsymbol{g}[T]$ that we call for simplicity G_1^T, where T can vary for every recording.

At recognition (testing) time, we can use the models to estimate the likelihood of a new sequence of observations G_1^T given each possible action, by means of the forward–backward inference algorithm. We can express this likelihood as $\mathcal{L}_{\text{HMM}}(G_1^T|A = a_k)$, where a_k is one of the possible actions. By normalizing the likelihoods, assuming that the gestures are equally likely *a priori*, we can obtain the posterior probability of the action given the sequence of observations as

$$P_{\text{HMM}}\big(A = a_k|G_1^T\big) = \frac{\mathcal{L}_{\text{HMM}}\big(G_1^T|A = a_k\big)}{\sum_h \mathcal{L}_{\text{HMM}}\big(G_1^T|A = a_h\big)}. \tag{1}$$

C. Combining the BN With Gesture HMMs

Once learned, the two models described above define two probability distributions over the relevant variables for the problem: 1) $P_{\text{BN}}(A, F, E, W)$ and 2) $P_{\text{HMM}}(A|G_1^T)$. The goal during inference is to merge the information provided by both models and estimate $P_{\text{comb}}(A, F, E, W|G_1^T)$, that is, the joint probability of all the affordance and word variables, given that we observe a certain action performed by the human.

To simplify the notation, we call $X = \{A, F, E, W\}$ the set of affordance and word variables $\{a, f_1, f_2, \ldots, e_1, e_2, \ldots, w_1, w_2, \ldots\}$. During inference, we have a (possibly empty) set of observed variables $X_{\text{obs}} \subseteq X$, and a set of variables $X_{\text{inf}} \subseteq X$ on which we wish to perform the inference. In order for the inference to be nontrivial, it must be $X_{\text{obs}} \cap X_{\text{inf}} = \varnothing$, that is, we should not observe any inference variable. According to the BN alone, the inference will compute the probability distribution of the inference variables X_{inf} given the observed variables X_{obs} by marginalizing over all the other (latent) variables $X_{\text{lat}} = X \backslash (X_{\text{obs}} \cup X_{\text{inf}})$, where $\backslash$ is the set difference operation

$$P_{\text{BN}}(X_{\text{inf}}|X_{\text{obs}}) = \sum_{X_{\text{lat}}} P_{\text{BN}}(X_{\text{inf}}, X_{\text{lat}}|X_{\text{obs}}).$$

If we want to combine the evidence brought by the BN with the evidence brought by the HMM, there are two cases that can occur:

1) the action variable is included among the inference variables: $A \in X_{\text{inf}}$; or
2) the action variable is not included among the inference variables: $A \in X_{\text{lat}}$.

Here, we are excluding the case where we observe the action directly ($A \in X_{\text{obs}}$) for two reasons. First, this would correspond to the robot performing it by itself, whereas we are interested in interpreting other people's actions, which is a necessary skill to engage in social collaboration with humans. Second, this would make the evidence on the gesture features G_1^T irrelevant, because in the model of Fig. 2 there is a tail-to-tail connection [47] from G_1^T to the rest of the variables through the action variable, which means that, given the action, all dependencies to the gesture features are dropped.

The two cases 1) and 2) enumerated above can be addressed separately when we do inference. In the first case, we call X'_{inf} the set of inference variables excluding the action A, that is, $X_{\text{inf}} = \{X'_{\text{inf}}, A\}$. We can write

$$\begin{aligned}
&P_{\text{comb}}\big(X_{\text{inf}}|X_{\text{obs}}, G_1^T\big) \\
&\quad = P_{\text{comb}}\big(A, X'_{\text{inf}}|X_{\text{obs}}, G_1^T\big) \\
&\quad = \sum_{X_{\text{lat}}} P_{\text{comb}}\big(A, X'_{\text{inf}}, X_{\text{lat}}|X_{\text{obs}}, G_1^T\big) \\
&\quad = \sum_{X_{\text{lat}}} \big[P_{\text{BN}}\big(A, X'_{\text{inf}}, X_{\text{lat}}|X_{\text{obs}}, G_1^T\big) \\
&\qquad\qquad P_{\text{HMM}}\big(A, X'_{\text{inf}}, X_{\text{lat}}|X_{\text{obs}}, G_1^T\big)\big] \\
&\quad = \left[\sum_{X_{\text{lat}}} P_{\text{BN}}\big(A, X'_{\text{inf}}, X_{\text{lat}}|X_{\text{obs}}\big)\right] P_{\text{HMM}}\big(A|G_1^T\big) \\
&\quad = P_{\text{BN}}(X_{\text{inf}}|X_{\text{obs}}) P_{\text{HMM}}\big(A|G_1^T\big).
\end{aligned} \tag{2}$$

This means that we can evaluate the two models independently, then multiply the distribution that we obtain from the BN (over all the possible value of the inference variables) by the HMM posterior for the corresponding value of the action.

In the second case, where the action is among the latent variables, we define, similarly, $X_{\text{lat}} = \{A, X'_{\text{lat}}\}$, and we have

$$\begin{aligned}
&P_{\text{comb}}\big(X_{\text{inf}}|X_{\text{obs}}, G_1^T\big) \\
&\quad = \sum_{\{A, X'_{\text{lat}}\}} P_{\text{comb}}\big(X_{\text{inf}}, A, X'_{\text{lat}}|X_{\text{obs}}, G_1^T\big) \\
&\quad = \sum_{\{A, X'_{\text{lat}}\}} \big[P_{\text{BN}}\big(X_{\text{inf}}, A, X'_{\text{lat}}|X_{\text{obs}}, G_1^T\big) \\
&\qquad\qquad P_{\text{HMM}}\big(X_{\text{inf}}, A, X'_{\text{lat}}|X_{\text{obs}}, G_1^T\big)\big] \\
&\quad = \sum_{\{A, X'_{\text{lat}}\}} \big[P_{\text{BN}}\big(X_{\text{inf}}, A, X'_{\text{lat}}|X_{\text{obs}}\big) P_{\text{HMM}}\big(A|G_1^T\big)\big] \\
&\quad = \sum_{A} \left[P_{\text{HMM}}\big(A|G_1^T\big) \sum_{X'_{\text{lat}}} P_{\text{BN}}\big(X_{\text{inf}}, A, X'_{\text{lat}}|X_{\text{obs}}\big)\right] \\
&\quad = \sum_{A} \big[P_{\text{HMM}}\big(A|G_1^T\big) P_{\text{BN}}(X_{\text{inf}}, A|X_{\text{obs}})\big].
\end{aligned} \tag{3}$$

This time, we first need to use the BN to do inference on the variables X_{inf} and A, and then we marginalize out the action variable A after having multiplied the probabilities by the HMM posterior.

D. Generation and Scoring of Verbal Descriptions

In order to illustrate the language capabilities of the model, rather than displaying the probability distribution of the words inferred by the model, we use the context-free grammar (CFG) described in the Appendix to generate written descriptions of the robot observations, on the basis of those probabilities. Note that this grammar is defined here with the only purpose of interpreting the probability distributions over the words. In the affordance–words model that we use, the speech recognizer is based on a free loop of words with uniform prior, and the Bayesian model relies on a bag-of-words assumption. No grammatical (syntactic) information about the spoken descriptions was, therefore, used during learning.

In the current study, by merging the affordance–words model and the gesture recognition model, we allow the robot to *reinterpret* the concepts it has learned in the self-centered phase, but we do not add any new words to the model. Consequently, the descriptions that the model generates when observing humans use the same words to describe the agent (see also Section V-E).

The textual descriptions are generated as follows: given some evidence X_{obs} that we provide to the model and some human observation features G_1^t extracted from frames 1 to t, we extract the generated word probabilities $P(w_i|X_{\text{obs}}, G_1^t)$. We generate N sentences randomly from the CFG using the `HSGen` tool from HTK [48]. Then, the sentences are rescored according to the log-likelihood of each word in the sentence, normalized by the length of the sentence

$$\text{score}\left(s_j|X_{\text{obs}}, G_1^t\right) = \frac{1}{L_j}\sum_{k=1}^{L_j} \log P\left(w_{jk}|X_{\text{obs}}, G_1^t\right) \tag{4}$$

where s_j is the jth sentence, L_j is the number of words in the sentence s_j, and w_{jk} is the kth word in the sentence s_j. Finally, an N-best list of possible descriptions is produced by sorting the scores.

IV. Experimental Settings

Our experiments consist on testing our method on a number of example scenarios that will be described in Section V. In this section, we provide experimental details and key assumptions of the method.

A. Affordance–Words Model

Table I presents a list of variables and the corresponding values used in the affordance–words model. Note that the name of the values of the affordance variables have been assigned by us arbitrarily to the clusters, for the sake of making the results more human-interpretable. However, the robot has no prior knowledge about the meaning of these clusters nor about their order, in case they correspond to ordered quantities. For extracting object features and effects from the sensory data, we assume that the robot possesses visual segmentation and geometric reasoning capabilities, meaning that it is able to segment the (potentially multiple) regions of interest corresponding to the physical objects of the world from the background (e.g., a planar surface such as a table) and to determine their positions.

TABLE I
Symbolic Variables of the Bayesian network (From [16]), With the Corresponding Discrete Values Obtained From Clustering During Robot Exploration of the Environment. We Call *Word Variables* the Booleans of the Last Row, Whereas We Call *Affordance Variables* All the Other Symbols

symbol	name: description	values
a	Action: motor action	grasp, tap, touch
f_1	Color: object color	blue, yellow, green1, green2
f_2	Size: object size	small, medium, big
f_3	Shape: object shape	sphere, box
e_1	ObjVel: object velocity	slow, medium, fast
e_2	HandVel: robot hand velocity	slow, fast
e_3	ObjHandVel: relative object–hand velocity	slow, medium, fast
e_4	Contact: object hand contact	short, long
w_1–w_{49}	presence of each word in the verbal description	true, false

We use the following notation in order to distinguish between the values of the affordance variables (all but the last row in Table I) and the words (last row in the table). Words and sentences are always enclosed in quotation marks. For example, "sphere" will refer to the spoken word, whereas sphere will refer to the value of the shape variable corresponding to the specific cluster. Similarly, "grasp" will correspond to a spoken word, whereas grasp corresponds to a value of the action variable.

There is no one-to-one correspondence between the values of the affordance variables and words. This was partly emerging from the natural variability that is inherent in the way humans describe situations in spoken words. It was also a design choice, because we wanted to prove that the model was not merely able to recover simple word–meaning associations, but was able to cope with more natural spoken utterances. Consequently, in the spoken descriptions, we have the following.

1) There are many synonyms for the same concept: for instance, cubic objects are called "box," "square," or "cube." Also, actions and effects are described using different tenses ["is grasping," "grasped," and "has (just) grasped"].
2) Different affordance variable values may have the same associated verbal description, e.g., two color clusters corresponding to different shades of green are both referred to as "green."
3) Finally, many affordance variable values have no direct description: for example, the object velocity and object–hand velocity (slow, medium, and fast), or the object–hand contact (short and long) are never described directly, and need to be inferred from the situation.

The affordance–words model does not account for the concepts of parts of speech, verb tenses or *temporal aspects* explicitly. For example, the words "is," "grasping," "has," "grasped," "just," and so on, are initially completely distinct and unrelated to the model, which has no prior information about what verbs, adjectives or nouns are, nor about similarity

between words. It is only through the association with the other robot observations that the model realizes that *grasping* has the same meaning as *grasped*. The following three phrases, which were used interchangeably in the experiments, are mapped to exactly the same meaning, after learning.

1) Is grasping.
2) Has grasped.
3) Grasped.

Note that the model *per se* would be fully capable to distinguish between those phrases, provided that they were used in different situations, which however was not the case in our experimental data.

B. Gesture Recognition

In this paper, we consider three independent, multiple-state HMMs, each of them trained to recognize one of the considered manipulation gestures of Fig. 1. The 3-D coordinates of the human limbs and torso used to extract the input to the gesture recognizer are obtained with a commodity depth sensor (Kinect).[4]

V. Results

In this section, we report the experimental findings obtained with our proposed model. Because it is based on Bayesian networks, the model can make inferences over any set of its variables X_{inf}, given any other set of observed variables X_{obs}. In particular, the model can do reasoning on the elements that constitute our computational concept of affordances, i.e., action, object features, and effects in Fig. 2. Furthermore, it can do reasoning over words. We present the following types of results.

1) Inferences over affordance variables (see Table I) in Sections V-A–V-C.
2) Predictions of word probabilities in Section V-D.
3) Verbal descriptions generated from the word probabilities of the previous point, according to a formal grammar. The descriptions, in turn, can be interpreted to observe the emergence of certain language phenomena (Sections V-E–V-G).

A. Action Recognition

In this experiment, we test the ability of our approach to recognize actions. Both affordance–words model and gesture recognition model can each perform inference of the Action variable individually: the former by using the variables of Table I, the latter by using human gesture features. We show how our method performs the inference over action in a joint way. This includes dealing with information with different degrees of confidence, or conflicting information.

The scene of Fig. 3 contains a small ball which, after the manipulative action, exhibits a low velocity. Based on the evidence, the affordance model gives the highest probability $P_{\text{BN}}(A|X_{\text{obs}})$ to the action *touch*, which usually does not result in any movement of the object. However, in this particular simulated situation, we assume that the action performed by the human was an (unsuccessful) *tap*, that is, a tap that does not result in any movement for the object. In the simulation we show the effect of augmenting the inference with information from a gesture recognizer, that is, computing $P_{\text{comb}}(A|X_{\text{obs}}, G_1^T)$. We analyze the effect of varying the degree of confidence of the classifier. We start from a uniform posterior $P_{\text{HMM}}(A|G_1^T)$, corresponding to a poor classifier, and gradually increase the probability of the correct action until it reaches 1. In this particular example, in order to win the belief of the affordance model, the action recognition needs to be very confident [$P_{\text{HMM}}(A = \text{tap}| G_1^T) > 0.81$].

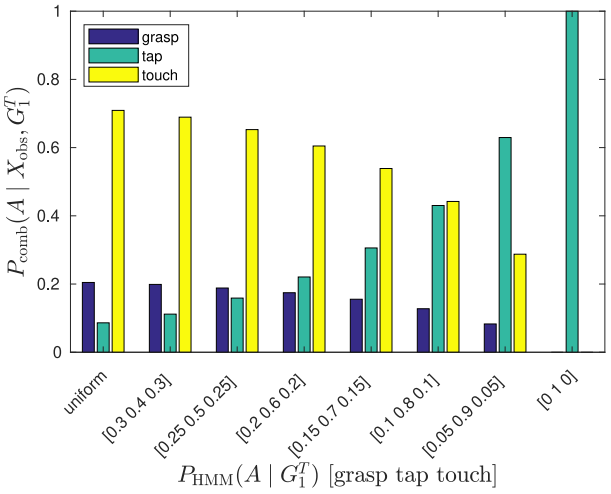


Fig. 3. Inference over action given the evidence X_{obs} = {Size = small, Shape = sphere, ObjVel = slow}, combined with different probabilistic soft evidence about the action.

B. Effect Prediction

We now show how our approach does inference over a different variable (instead of the action one which is common between affordance–words model and gesture model), i.e., how it predicts the value of the object velocity effect variable. We will do this by using different degrees of probabilistic confidence about the action, and analyzing the outcome in terms of velocity prediction. This experiment exposes that *all* the variables of Table I jointly link robot and human, not only the Action variable, for the reasons expressed in Section III.

Fig. 4 shows the considered inference in two cases: 1) when the prior information says that the shape is spherical [see Fig. 4(a)] and 2) when it is cubic [see Fig. 4(b)].

The leftmost distribution in both figures shows the prediction of object velocity from the affordance–words model alone, without any additional information. When the shape is spherical, the model is not sure about the velocity, whereas if the shape is cubic, the model does not expect high velocities. If we add clear evidence on the action *touch* from the

[4]Currently, our gesture recognition algorithm relies on human skeleton tracking software from a depth stream. In our experience, the hand tracking is not reliable in the presence of a tabletop (i.e., partially occluded human) as in Fig. 1, so we record the same gestures twice, with and without the table: the latter is used for ensuring the robustness of the estimated hand coordinate, the former is used throughout the rest of our model and experiments. We plan to overcome this limitation in future work.

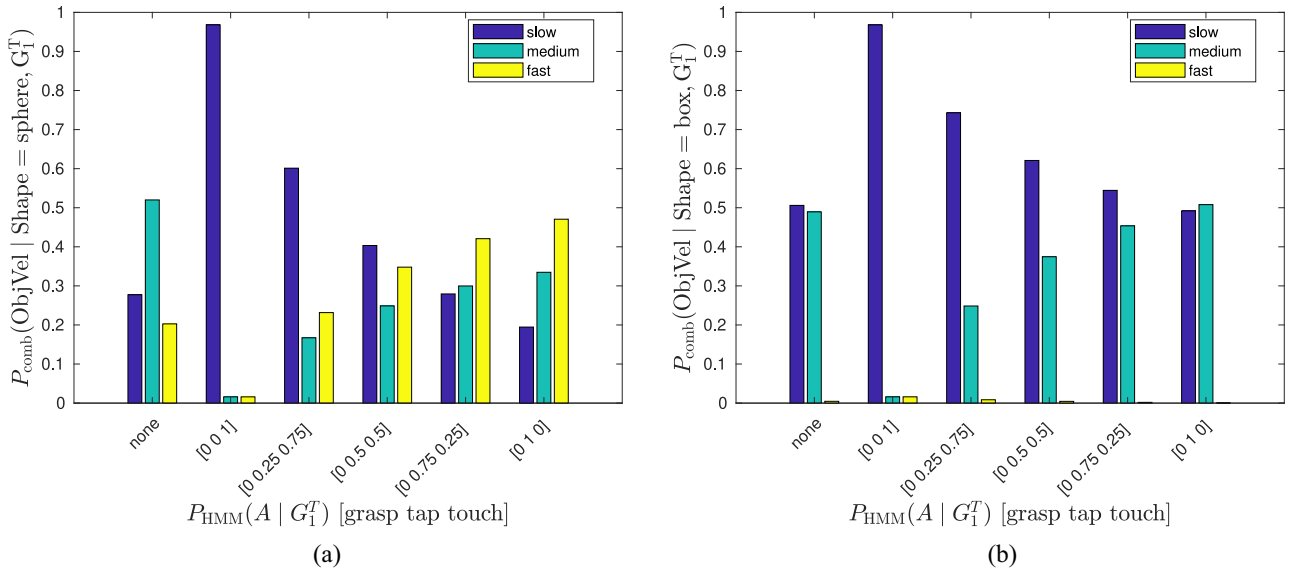


Fig. 4. Inference over the object velocity effect of different objects, when given probabilistic soft evidence about the action. Predictions with (a) sphere object and (b) box object.

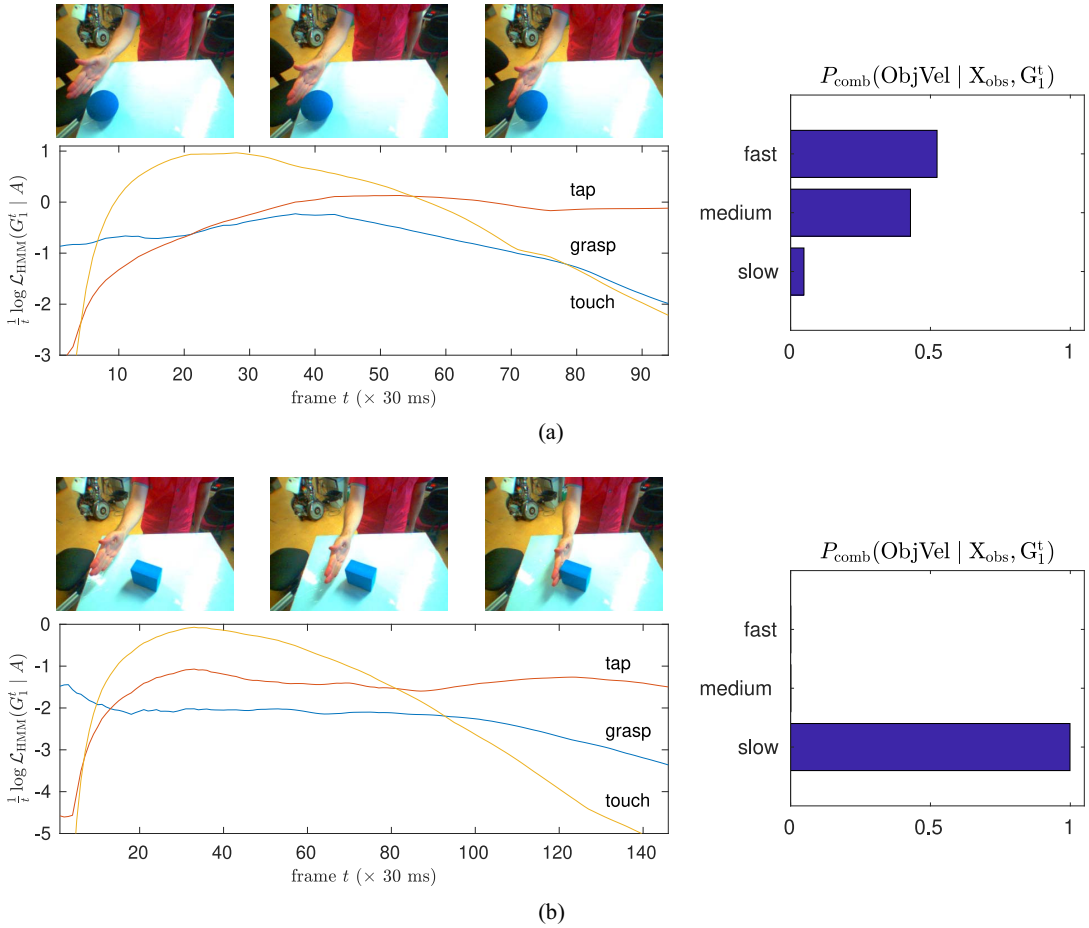


Fig. 5. Object velocity effect anticipation before impact. The evidence from the gesture recognizer (left) is fed into the affordance–words model before the end of the execution. The combined model predicts the effect (right) and describes it in words. (a) Action performed on small sphere. Description: "the robot pushed the ball and the ball moves." (b) Action performed on big box. Description: "the robot is pushing the big square but the box is inert."

action recognition model, suddenly the combined model predicts slow velocities in both cases, as expected. However, if the action recognition evidence is gradually changed from *touch* to *tap*, the predictions of the model depend on the shape of the object. Higher velocities are expected for spherical objects that can roll, compared to cubic objects.

C. Effect Anticipation

Since the gesture recognition method interprets sequences of human motions, we can test this predictive ability of the complete model when we observe an incomplete action. Fig. 5 shows an example of this where we reason about the expected object velocity caused by a tap action. Fig. 5(a) shows the action performed on a spherical object, whereas Fig. 5(b) on a cubic one. The graphs on the left side show the time evolution of the evidence $P_{\text{HMM}}(A|G_1^t)$ from the gesture recognition model. In order to make the variations emerge more clearly, instead of the posterior, we show $(1/t)\log\mathcal{L}_{\text{HMM}}(G_1^t|A)$: the log-likelihood normalized by the length of the sequence. Note how, in both cases, the correct action is recognized by the model given enough evidence, although the observation sequence is not complete. The right side of the plot shows the prediction of the object velocity, given the incomplete observation of the action and the object properties. The model correctly predicts that the sphere will probably move but the box is unlikely do so. Finally, the captions in the figure also show the verbal description (see Section III-D) generated by feeding the probability distribution of the words estimated by the model given the evidence into the context-free grammar.

D. Prediction of Word Probabilities

Our model permits to make predictions over the word variables associated to affordance evidence. In Fig. 6, we show the variation in word occurrence probabilities between two cases.

1) When the robot's prior knowledge evidence consists of information about object features and effects only: {Size=big, Shape=sphere, ObjVel=fast}.
2) When the evidence corresponds to the one of the previous point, with the addition of the *tap* action observed from the gesture recognizer (hard evidence).

In this result, we notice two facts. First, the probabilities of words related to tapping and pushing increase when a tapping action evidence from gesture recognition is introduced; conversely, the probabilities of other action words (touching and poking) decreases. Second, the probability of the word "rolling" (which is an effect of an action onto an object) also increases when the tap action evidence is entered.

E. Verbal Descriptions and Choice of Synonyms

By generating and scoring natural language descriptions of what the robot observes (see Section III-D), we can provide evidence to the model and interpret the verbal results. Recall that, with our method, we do not add new words to the model when we observe the human performing actions. Rather, the human-readable descriptions that we generate are based on the same words that were present in the self-centered learning phase. In this phase, the verbal descriptions described the agent of the observed actions is either "the robot," "he," or "Baltazar" (the name of the robot). Consequently, the affordance–words model learned by the robot includes those words as the subject of the action.

As an example, by providing the evidence {Color=yellow, Size=big, Shape=sphere, ObjVel=fast} to the model, we obtain the sentences reported in Table II. The higher the score, the better. In many of these sentences, we note that: 1) the correct verb related to the tap action is generated (in the initial evidence, no action information was present, only object features and effects information were) and 2) the object term "ball" or synonyms thereof (e.g., sphere) are used coherently, both in the first part of the sentence describing the action and in the second part describing the effect. The fact that different synonyms may be used in the same sentence is simply a consequence of the random generation of sentences, described in Section III-D, and of the fact that usually synonyms are assigned similar (but not necessarily equal) probabilities by the model, given the same evidence.

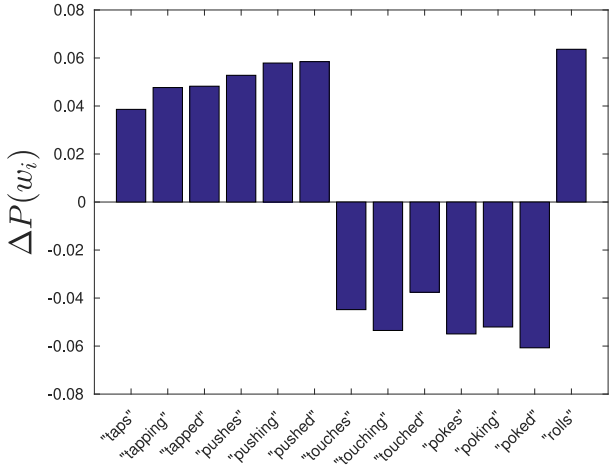


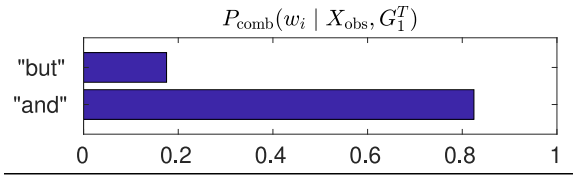
Fig. 6. Variation of word occurrence probabilities: $\Delta P(w_i) = P_{\text{comb}}(w_i|X_{\text{obs}}, \text{Action=tap}) - P_{\text{BN}}(w_i|X_{\text{obs}})$, where $X_{\text{obs}} =$ {Size=big, Shape=sphere, ObjVel=fast}. This variation corresponds to the difference of word probability when we add the tap action evidence (obtained from gesture recognition) to the initial evidence about object features and effects. We have omitted words for which no significant variation was observed.

TABLE II
TEN-BEST LIST OF SENTENCES GENERATED FROM THE EVIDENCE $X_{\text{OBS}} =$ {COLOR=YELLOW, SIZE=BIG, SHAPE=SPHERE, OBJVEL=FAST}

sentence	score
"the robot pushed the ball and the ball moves"	-0.54322
"the robot tapped the sphere and the sphere moves"	-0.5605
"he is pushing the sphere and the sphere moves"	-0.57731
"the robot is tapping the yellow ball and the big yellow sphere is moving"	-0.57932
"he pushed the yellow ball and the sphere is rolling"	-0.58853
"the robot is poking the ball and the sphere is rolling"	-0.58998
"he is pushing the ball and the yellow ball moves"	-0.59728
"he pushes the sphere and the ball is moving"	-0.60528
"he is tapping the yellow ball and the ball is moving"	-0.60675
"the robot pokes the sphere and the ball is rolling"	-0.60694

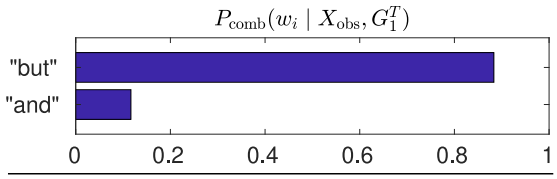
F. Language Phenomenon: Choice of Correct Conjunction

The manipulation experiments that we consider have the following structure: an agent (human or robot) performs a physical action onto an object with certain properties, and this object will produce a certain physical effect as a result. For example, a touch action on an object yields no physical movement, but a tap does (especially if the object is spherical). In the language description associated to an experiment, it makes sense to measure the conjunction chosen by the model



sentence	score
"the robot is picking the sphere **and** the sphere is moving"	−0.59328
"the robot grasps the sphere **and** the ball is moving"	−0.59507
"the robot is picking the sphere **and** the sphere is rising"	−0.60882
"the robot grasped the sphere **and** the sphere is rising"	−0.61842
"the robot picked the ball **and** the ball is rising"	−0.64052
"baltazar grasps the sphere **and** the sphere is moving"	−0.66182
"the robot has grasped the ball **and** the ball is rising"	−0.66398
"the robot picked the ball **and** the green ball is moving"	−0.67134
"baltazar grasped the sphere **and** the ball is moving"	−0.67283
"baltazar is grasping the ball **and** the sphere is rising"	−0.6787

(a)



sentence	score
"the robot is picking the cube **but** the square is still"	−0.52575
"the robot is grasping the sphere **but** the box is inert"	−0.55
"the robot is grasping the square **but** the sphere is still"	−0.55388
"the robot grasped the square **but** the cube is inert"	−0.55608
"baltazar is grasping the square **but** the square is inert"	−0.5571
"the robot is grasping the cube **but** the ball is inert"	−0.56011
"the robot picks the box **but** the square is inert"	−0.56397
"baltazar is picking the square **but** the square is still"	−0.56402
"he is grasping the square **but** the cube is inert"	−0.56815
"the robot grasps the square **but** the sphere is inert"	−0.57417

(b)

Fig. 7. Ten-best list of sentences generated given two different sets of evidence. In (a), the model interprets the object movement as indicating a successful grasp and uses the conjunction "and." In (b), the slow movement is interpreted as no movement at all and, therefore, as an unsuccessful grasp: for that reason, the conjunction "but" is used. (a) Evidence: X_{obs} ={Action=grasp, ObjVel=medium}. (b) Evidence: X_{obs} ={Action=grasp, ObjVel=slow}.

(a)

(b)

(c)

(d)

Fig. 8. Example of descriptions generated by the model. (a) The robot is grasping the box and the green box is moving." (b) The robot is poking the green square and the cube is inert." (c) The robot picked the ball and the green ball is moving." (d) Baltazar is poking the green sphere and the sphere is still."

given specific evidence. In particular, it would be desirable to separate two kinds of behaviors: one in which the action and effect are coherent (expected conjunction: "and"), and the other one in which they are contradictory (but).

Fig. 7 shows an example of this behavior of the model. We give the same action value *grasp* to the model as evidence, but two different values for the final object velocity. When the object velocity is medium [Fig. 7(a)], the model interprets this as a successful grasp and uses the conjunction "and" to separate the description of the action from the description of the effect. When the object velocity is slow (in the clustering procedure, the velocity was most often zero in those cases), the model predicts that this is an unsuccessful grasp and uses the conjunction "but," instead.

G. Language Phenomenon: Description of Object Features

In Fig. 8, we show examples of verbal descriptions generated by the model given different values of observed evidence.

1) X_{obs} = {Action=grasp, Color=green1, Shape=box} [Fig. 8(a)].
2) X_{obs} = {Action=touch, Color=green1, Shape=box} [Fig. 8(b)].
3) X_{obs} = {Action=grasp, Color=green2, Shape=sphere} [Fig. 8(c)].
4) X_{obs} = {Action=touch, Color=green2, Shape=sphere} [Fig. 8(d)].

Note that the box object in the two first examples has a dark shade of green (value of color affordance variable of Table I clustered as: green1), whereas the spherical one in the two last examples has a lighter shade (color value: green2). However, the verbal descriptions reported in Fig. 8 all use the adjective "green." This behavior emerges from fact that the robot develops its perceptual symbols (clusters) in an early phase, and only subsequently associates them with the human vocabulary. We believe that this phenomenon is practical and potentially useful (i.e., the possibility that a low-level fine-grained robot representation can be abstracted into a high-level language description, which bundles the two shades of green under the same word).

VI. Conclusion

We presented a model that allows a robot to interpret and describe the actions of external agents, by reusing the knowledge previously acquired in an ego-centric manner. In a developmental setting, the robot first learns the link between words and object affordances by exploring its environment. Then, it uses this information to learn to classify the gestures and actions of another agent. Finally, by fusing the information from the two probabilistic models, in our experiments we show that the robot can reason over affordances and words when observing the other agent; this can also be leveraged to do early action recognition (see Section V-C). Although the

complete model only estimates probabilities of single words given the evidence, we showed that feeding these probabilities into a predefined grammar produces human-interpretable sentences that correctly describe the situation. We also highlighted some interesting language-related properties of the model, such as: congruent/incongruent conjunctions, choice of appropriate synonym words, and describing object features with general words.

Our demonstrations are based on a restricted scenario (see Section IV), i.e., one human and one robot manipulating simple objects on a shared table, a predefined number of motor actions and effects, and a vocabulary of approximately 50 words to describe the experiments verbally. However, one of the main strengths of this paper is that it spans different fields, such as robot learning, language grounding, and object affordances. We also work with real robotic data, as opposed to learning images-to-text mappings (as in many works in computer vision) or using robot simulations (as in many works in robotics).

In terms of *scalability*, note that our BN model can learn both dependency structure and parameters of the model from observations. The method that estimates the dependency structure, in particular, is sensitive to biases in the data. Consequently, in order to avoid misconceptions, the robot needs to explore any possible situation that may occur. For example, if the robot only observes blue spheres rolling, it might infer that it is the color that makes the object roll, rather than its shape. In order to scale the method to a larger number of concepts, it would be necessary to scale the amount of data considerably, similarly to what is typically done in deep learning. In models of developmental robotics, where this is neither practically feasible, nor desirable, we would need to devise methods that can generalize more efficiently from very few observations.

As future work, it would be useful to investigate how the model can extract syntactic information from the observed data autonomously, thus relaxing the bag-of-words assumption in the current model. Another line of research would be to study how the model can guide the discovery of new acoustic patterns (e.g., [49]–[51]), and how to incorporate the newly discovered symbols into our affordance–words model. This would release our current assumption of a predefined set of words.

APPENDIX
GRAMMAR DEFINITION

Below, we provide the grammar definition used to generate verbal descriptions from the probability distribution over words estimated by the model. Note, however, that no grammar was used during the learning phase: the speech recognizer used as a frontend to the spoken descriptions is based on a loop of words with no grammar, and the affordance–words model is based on a bag-of-words assumption, where only the presence or absence of each word in the description is considered. The symbol `.|.` represents alternative items, while the symbol `[.]` optional items. Nonterminal symbols are given between `<.>` in italics, while words (terminal symbols) are given in plain text and font: thus, the full set of words is given by all the plain text words

⟨ *sentence* ⟩ ::= ⟨ *agent* ⟩ ⟨ *action* ⟩ ⟨ *object* ⟩ ⟨ *conjunction* ⟩ ⟨ *object* ⟩ ⟨ *effect* ⟩

⟨ *agent* ⟩ ::= the robot | he | baltazar

⟨ *action* ⟩ ::= ⟨ *touch* ⟩ | ⟨ *poke* ⟩ | ⟨ *tap* ⟩ | ⟨ *push* ⟩ | ⟨ *grasp* ⟩ | ⟨ *pick* ⟩

⟨ *touch* ⟩ ::= touches | [has] [just] touched | is touching

⟨ *poke* ⟩ ::= pokes | [has] [just] poked | is poking

⟨ *tap* ⟩ ::= taps | [has] [just] tapped | is tapping

⟨ *push* ⟩ ::= pushes | [has] [just] pushed | is pushing

⟨ *grasp* ⟩ ::= grasps | [has] [just] grasped | is grasping

⟨ *pick* ⟩ ::= picks | [has] [just] picked | is picking

⟨ *object* ⟩ ::= the [⟨ *size* ⟩] [⟨ *color* ⟩] ⟨ *shape* ⟩

⟨ *size* ⟩ ::= big | small

⟨ *color* ⟩ ::= green | yellow | blue

⟨ *shape* ⟩ ::= sphere | ball | cube | box | square

⟨ *conjunction* ⟩ ::= and | but

⟨ *effect* ⟩ ::= ⟨ *inertmove* ⟩ | ⟨ *slideroll* ⟩ | ⟨ *fallrise* ⟩

⟨ *inertmove* ⟩ ::= is inert | is still | moves | is moving

⟨ *slideroll* ⟩ ::= slides | is sliding | rolls | is rolling

⟨ *fallrise* ⟩ ::= rises | is rising | falls | is falling.

REFERENCES

[1] V. W. Turner, *Dramas, Fields, and Metaphors: Symbolic Action in Human Society*. Ithaca, NY, USA: Cornell Univ. Press, 1975.

[2] C. A. Brownell, G. B. Ramani, and S. Zerwas, "Becoming a social partner with peers: Cooperation and social understanding in one-and two-year-olds," *Child Develop.*, vol. 77, no. 4, pp. 803–821, 2006.

[3] A. P. Melis and D. Semmann, "How is human cooperation different?" *Philosoph. Trans. Roy. Soc. London B Biol. Sci.*, vol. 365, no. 1553, pp. 2663–2674, 2010.

[4] N. Ramnani and R. C. Miall, "A system in the human brain for predicting the actions of others," *Nat. Neurosci.*, vol. 7, no. 1, pp. 85–90, 2004.

[5] C. L. Breazeal, *Designing Sociable Robots*. Cambridge, MA, USA: MIT Press, 2002.

[6] N. Iwahashi, "Robots that learn language: A developmental approach to situated human–robot conversations," in *Human Robot Interaction*. Rijeka, Croatia: InTech, 2007, ch. 5. doi: 10.5772/5188.

[7] L. Steels, "Evolving grounded communication for robots," *Trends Cogn. Sci.*, vol. 7, no. 7, pp. 308–312, 2003.

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

[9] A. Cangelosi and M. Schlesinger, *Developmental Robotics: From Babies to Robots*. Cambridge, MA, USA: MIT Press, 2015.

[10] S. Gallagher and A. N. Meltzoff, "The earliest sense of self and others: Merleau–Ponty and recent developmental studies," *Philosoph. Psychol.*, vol. 9, no. 2, pp. 211–233, 1996.

[11] G. Rizzolatti, L. Fogassi, and V. Gallese, "Neurophysiological mechanisms underlying the understanding and imitation of action," *Nat. Rev. Neurosci.*, vol. 2, no. 9, pp. 661–670, 2001.

[12] J. Decety and J. A. Sommerville, "Shared representations between self and other: A social cognitive neuroscience view," *Trends Cogn. Sci.*, vol. 7, no. 12, pp. 527–533, 2003.

[13] D. K. Symons, "Mental state discourse, theory of mind, and the internalization of self–other understanding," *Develop. Rev.*, vol. 24, no. 2, pp. 159–188, 2004.

[14] F. Ribordy, A. Jabès, P. B. Lavenex, and P. Lavenex, "Development of allocentric spatial memory abilities in children from 18 months to 5 years of age," *Cogn. Psychol.*, vol. 66, no. 1, pp. 1–29, 2013.

[15] G. Saponaro, L. Jamone, A. Bernardino, and G. Salvi, "Interactive robot learning of gestures, language and affordances," in *Proc. Int. Workshop Grounding Lang. Understand.*, 2017, pp. 83–87.
[16] G. Salvi, L. Montesano, A. Bernardino, and J. Santos-Victor, "Language bootstrapping: Learning word meanings from perception–action association," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 3, pp. 660–671, Jun. 2012.
[17] G. Saponaro, G. Salvi, and A. Bernardino, "Robot anticipation of human intentions through continuous gesture recognition," in *Proc. Int. Conf. Collaborat. Technol. Syst.*, 2013, pp. 218–225.
[18] U. Müller and J. I. M. Carpendale, "The role of social interaction in Piaget's theory: Language for social cooperation and social cooperation for language," *New Ideas Psychol.*, vol. 18, nos. 2–3, pp. 139–156, 2000.
[19] S. Rojas-Drummond and N. Mercer, "Scaffolding the development of effective collaboration and learning," *Int. J. Edu. Res.*, vol. 39, nos. 1–2, pp. 99–111, 2003.
[20] G. Schillaci, B. Lara, and V. V. Hafner, "Internal simulations for behaviour selection and recognition," in *Proc. Int. Workshop Human Behav. Understand.*, 2012, pp. 148–160.
[21] E. A. Billing, H. Svensson, R. Lowe, and T. Ziemke, "Finding your way from the bed to the kitchen: Reenacting and recombining sensorimotor episodes learned from human demonstration," *Front. Robot. AI*, vol. 3, no. 9, 2019. doi: 10.3389/frobt.2016.00009.
[22] S. M. Aglioti, P. Cesari, M. Romani, and C. Urgesi, "Action anticipation and motor resonance in elite basketball players," *Nat. Neurosci.*, vol. 11, no. 9, pp. 1109–1116, 2008.
[23] G. Knoblich and R. Flach, "Predicting the effects of actions: Interactions of perception and action," *Psychol. Sci.*, vol. 12, no. 6, pp. 467–472, 2001.
[24] V. Gazzola, G. Rizzolatti, B. Wicker, and C. Keysers, "The anthropomorphic brain: The mirror neuron system responds to human and robotic actions," *Neuroimage*, vol. 35, no. 4, pp. 1674–1684, 2007.
[25] M. Lopes, F. S. Melo, B. Kenward, and J. Santos-Victor, "A computational model of social-learning mechanisms," *Adapt. Behav.*, vol. 17, no. 6, pp. 467–483, 2009.
[26] E. Ugur, Y. Nagai, H. Celikkanat, and E. Oztop, "Parental scaffolding as a bootstrapping mechanism for learning grasp affordances and imitation skills," *Robotica*, vol. 33, no. 5, pp. 1163–1180, 2015.
[27] E. Ugur, Y. Nagai, E. Sahin, and E. Oztop, "Staged development of robot skills: Behavior formation, affordance learning and imitation with motionese," *IEEE Trans. Auton. Mental Develop.*, vol. 7, no. 2, pp. 119–139, Jun. 2015.
[28] C. L. Nehaniv and K. Dautenhahn, "The correspondence problem," in *Imitation in Animals and Artifacts*. Cambridge, MA, USA: MIT Press, 2002, ch. 2.
[29] S. Kim, Z. Yu, and M. Lee, "Understanding human intention by connecting perception and action learning in artificial agents," *Neural Netw.*, vol. 92, pp. 29–38, Aug. 2017.
[30] C. Szegedy *et al.*, "Intriguing properties of neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2014, pp. 1–10. [Online]. Available: https://iclr.cc/archive/2014/conference-proceedings/
[31] A. Santoro *et al.*, "A simple neural network module for relational reasoning," in *Proc. Conf. Neural Inf. Process. Syst.*, 2017, pp. 4974–4983.
[32] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from RGB-D videos," *Int. J. Robot. Res.*, vol. 32, no. 8, pp. 951–970, 2013.
[33] S. Harnad, "The symbol grounding problem," *Physica D Nonlin. Phenomena*, vol. 42, nos. 1–3, pp. 335–346, 1990.
[34] C. Matuszek, L. Bo, L. Zettlemoyer, and D. Fox, "Learning from unscripted deictic gesture and language for human–robot interactions," in *Proc. AAAI Conf. Artif. Intell.*, 2014, pp. 2556–2563.
[35] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building machines that learn and think like people," *Behav. Brain Sci.*, vol. 40, p. e253, 2017. doi: 10.1017/S0140525X16001837.
[36] Q. Gao, S. Yang, J. Chai, and L. Vanderwende, "What action causes this? Towards naive physical action–effect prediction," in *Proc. Annu. Meeting Assoc. Comput. Linguist.*, 2018, pp. 934–945.
[37] J. J. Gibson, *The Ecological Approach to Visual Perception: Classic Edition*. New York, NY, USA: Psychol. Press, 2014.
[38] L. Montesano, M. Lopes, A. Bernardino, and J. Santos-Victor, "Learning object affordances: From sensory–motor maps to imitation," *IEEE Trans. Robot.*, vol. 24, no. 1, pp. 15–26, Feb. 2008.
[39] L. Jamone *et al.*, "Affordances in psychology, neuroscience, and robotics: A survey," *IEEE Trans. Cogn. Devel. Syst.*, vol. 10, no. 1, pp. 4–25, Mar. 2018.
[40] C. Maestre, G. Mukhtar, C. Gonzales, and S. Doncieux, "Iterative affordance learning with adaptive action generation," in *Proc. IEEE Int. Conf. Develop. Learn. Epigenetic Robot.*, 2017, pp. 368–375.
[41] P. Zech *et al.*, "Computational models of affordance in robotics: A taxonomy and systematic classification," *Adapt. Behav.*, vol. 25, no. 5, pp. 235–271, 2017.
[42] T. Nakamura, T. Nagai, and N. Iwahashi, "Grounding of word meanings in multimodal concepts using LDA," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2009, pp. 3943–3948.
[43] T. Araki *et al.*, "Online learning of concepts and words using multimodal LDA and hierarchical Pitman–Yor language model," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2012, pp. 1623–1630.
[44] A. F. Morse and A. Cangelosi, "Why are there developmental stages in language learning? A developmental robotics model of language development," *Cogn. Sci.*, vol. 41, pp. 32–51, Feb. 2017.
[45] F. Stramandinoli, V. Tikhanoff, U. Pattacini, and F. Nori, "Grounding speech utterances in robotics affordances: An embodied statistical language model," in *Proc. IEEE Int. Conf. Develop. Learn. Epigenetic Robot.*, 2016, pp. 79–86.
[46] A. Antunes, L. Jamone, G. Saponaro, A. Bernardino, and R. Ventura, "From human instructions to robot actions: Formulation of goals, affordances and probabilistic planning," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2016, pp. 5449–5454.
[47] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Burlington, MA, USA: Morgan Kaufmann, 2014.
[48] S. Young *et al.*, *The HTK Book (for HTK Version 3.4)*, Dept. Eng., Cambridge Univ., Cambridge, U.K., 2006.
[49] A. F. Myrman and G. Salvi, "Partitioning of posteriorgrams using siamese models for unsupervised acoustic modelling," in *Proc. Int. Workshop Grounding Lang. Understand.*, 2017, pp. 27–31.
[50] N. Vanhainen and G. Salvi, "Pattern discovery in continuous speech using block diagonal infinite HMM," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2014, pp. 3719–3723.
[51] N. Vanhainen and G. Salvi, "Word discovery with beta process factor analysis," in *Proc. Int. Conf. Speech Commun. Technol.*, 2012, pp. 798–802.

Giovanni Saponaro (S'11) received the B.Sc. degree in computer engineering and the M.Sc. degree (honors) in computer engineering—artificial intelligence systems from the Sapienza University of Rome, Rome, Italy, in 2005 and 2009, respectively. He is currently working toward the Ph.D. degree at the Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal.

He is a member of the Computer and Robot Vision Laboratory, Institute for Systems and Robotics, Instituto Superior Técnico, Universidade de Lisboa. He has participated in the international research project POETICON++, together with linguists, computer vision experts, neuroscientists, and roboticists. He has published over ten papers in the diverse areas of cognitive systems, developmental robotics, visual perception of objects and of human body gestures for action recognition. His current research interests include visual scene understanding and robot decision algorithms that support human–robot interaction with highly advanced humanoid robots such as the iCub, and in the presence of uncertainty.

Lorenzo Jamone (M'13) received the M.S. degree (honors) in computer engineering from the University of Genoa, Genoa, Italy, in 2006, and the Ph.D. degree in humanoid technologies from the University of Genoa, Genoa, Italy, and the Italian Institute of Technology, Genoa, in 2010.

He is a Lecturer of Robotics at the Queen Mary University of London, London, U.K. He was an Associate Researcher at the Takanishi Laboratory, Waseda University, Tokyo, Japan, from 2010 to 2012, and the Computer and Robot Vision Laboratory, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal, from 2012 to 2016. He has over 60 publications with an *H*-index of 16. His current research interests include cognitive humanoid robots, sensorimotor learning and control, robotic manipulation, and force and tactile sensing.

Alexandre Bernardino (SM'04) received the Ph.D. degree in electrical and computer engineering, in 2004.

He is an Associate Professor at the Department of Electrical and Computer Engineering and a Senior Researcher at the Computer and Robot Vision Laboratory, Institute for Systems and Robotics, Instituto Superior Técnico, Faculty of Engineering, Universidade de Lisboa, Lisbon, Portugal. He has graduated ten Ph.D. students and over 40 M.Sc. students. He has participated in several national and international research projects as a principal investigator and a technical manager. He has published over 40 research papers in peer-reviewed journals and over 100 papers on peer-reviewed conferences in the fields of robotics, vision, and cognitive systems. His current research interests include application of computer vision, machine learning, cognitive science, and control theory to advanced robotics and automation systems.

Dr. Bernardino is an Associate Editor of the journal *Frontiers in Robotics and AI* and major robotics conferences, such as ICRA and IROS. He was a Co-Supervisor of the Ph.D. thesis that won the IBM Prize 2014, and the Supervisor of the Best Robotics Portuguese M.Sc. thesis award of 2012. He is currently the Chair of the IEEE Portugal Robotics and Automation Chapter.

Giampiero Salvi received the M.Sc. degree in electrical engineering from the Sapienza University of Rome, Rome, Italy, and the Ph.D. degree in computer science from the KTH Royal Institute of Technology, Stockholm, Sweden.

He was a Postdoctoral Fellow at the Institute of Systems and Robotics, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal. He is currently an Associate Professor of Machine Learning and the Director of the Masters Programme in Machine Learning at the KTH Royal Institute of Technology. His current research interests include machine learning, speech technology, and cognitive systems.