International Workshop on Grounding Language Understanding (GLU), INTERSPEECH 2017
Stockholm, Sweden, 25 August 2017

Interactive Robot Learning of Gestures, Language and Affordances

Giovanni Saponaro[1], *Lorenzo Jamone*[2,1], *Alexandre Bernardino*[1], *Giampiero Salvi*[3]

[1]Institute for Systems and Robotics
Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal
[2]ARQ (Advanced Robotics at Queen Mary)
School of Electronic Engineering and Computer Science, Queen Mary University of London, UK
[3]KTH Royal Institute of Technology, Stockholm, Sweden

gsaponaro@isr.tecnico.ulisboa.pt, l.jamone@qmul.ac.uk, alex@isr.tecnico.ulisboa.pt, giampi@kth.se

Abstract

A growing field in robotics and Artificial Intelligence (AI) research is human–robot collaboration, whose target is to enable effective teamwork between humans and robots. However, in many situations human teams are still superior to human–robot teams, primarily because human teams can easily agree on a common goal with language, and the individual members observe each other effectively, leveraging their shared motor repertoire and sensorimotor resources. This paper shows that for cognitive robots it is possible, and indeed fruitful, to combine knowledge acquired from interacting with elements of the environment (affordance exploration) with the probabilistic observation of another agent's actions.

We propose a model that unites (i) learning robot affordances and word descriptions with (ii) statistical recognition of human gestures with vision sensors. We discuss theoretical motivations, possible implementations, and we show initial results which highlight that, after having acquired knowledge of its surrounding environment, a humanoid robot can generalize this knowledge to the case when it observes another agent (human partner) performing the same motor actions previously executed during training.

Index Terms: cognitive robotics, gesture recognition, object affordances

1. Introduction

Robotics is progressing fast, with a steady and systematic shift from the industrial domain to domestic, public and leisure environments [1, ch. 65, Domestic Robotics]. Application areas that are particularly relevant and being researched by the scientific community include: robots for people's health and active aging, mobility, advanced manufacturing (Industry 4.0). In short, all domains that require direct and effective human–robot interaction and communication (including language and gestures [2]).

However, robots have not reached the level of performance that would enable them to work with humans in routine activities in a flexible and adaptive way, for example in the presence of sensor noise, or unexpected events not previously seen during the training or learning phase. One of the reasons to explain this performance gap between human–human teamwork and a human–robot teamwork is in the collaboration aspect, i. e., whether the members of a team understand one another. Humans have the ability of working successfully in groups. They can agree on common goals (e. g., through verbal and non-verbal communication), work towards the execution of these goals in a coordinated way, and understand each other's physical actions (e. g., body gestures) towards the realization of the final target. Human team coordination and mutual understanding is effective [3] because of (i) the capacity to *adapt* to unforeseen events in the environment, and re-plan one's actions in real time if necessary, and (ii) a common motor repertoire and action model, which permits us to understand a partner's physical actions and manifested intentions as if they were our own [4].

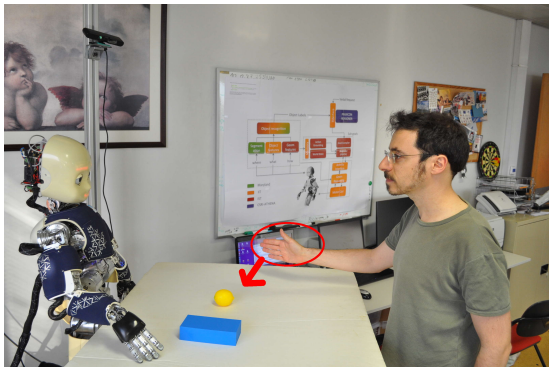


Figure 1: *Experimental setup, consisting of an iCub humanoid robot and a human user performing a manipulation gesture on a shared table with different objects on top. The depth sensor in the top-left corner is used to extract human hand coordinates for gesture recognition. Depending on the gesture and on the target object, the resulting effect will differ.*

In neuroscience research, visuomotor neurons (i. e., neurons that are activated by visual stimuli) have been a subject of ample study [5]. Mirror neurons are one class of such neurons that responds to action and object interaction, both when the agent acts and when it observes the same action performed by others, hence the name "mirror".

This work takes inspiration from the theory of mirror neurons, and contributes towards using it on humanoid and cognitive robots. We show that a robot can first acquire knowledge by sensing and self-exploring its surrounding environment (e. g., by interacting with available objects and building up an affordance representation of the interactions and their outcomes) and, as a result, the robot is capable of generalizing its acquired knowledge while observing another agent (e. g., a human person) who performs similar physical actions to the ones executed during prior robot training. Fig. 1 shows the experimental setup.

2. Related Work

A large and growing body of research is directed towards having robots learn new cognitive skills, or improving their capabilities, by interacting autonomously with their surrounding environment. In particular, robots operating in an unstructured scenario may understand available opportunities conditioned on their body, perception and sensorimotor experiences: the intersection of these elements gives rise to object affordances (action possibilities), as they are called in psychology [6]. The usefulness of affordances in cognitive robotics is in the fact that they capture essential properties of environment objects in terms of the actions that a robot is able to perform with them [7, 8]. Some authors have suggested an alternative computational model called Object–Action Complexes (OACs) [9], which links low-level sensorimotor knowledge with high-level symbolic reasoning hierarchically in autonomous robots.

In addition, several works have demonstrated how combining robot affordance learning with language grounding can provide cognitive robots with new and useful skills, such as learning the association of spoken words with sensorimotor experience [10, 11] or sensorimotor representations [12], learning tool use capabilities [13, 14], and carrying out complex manipulation tasks expressed in natural language instructions which require planning and reasoning [15].

In [10], a joint model is proposed to learn robot affordances (i. e., relationships between actions, objects and resulting effects) together with word meanings. The data contains robot manipulation experiments, each of them associated with a number of alternative verbal descriptions uttered by two speakers for a total of 1270 recordings. That framework assumes that the robot action is known a priori during the training phase (e. g., the information "grasping" during a grasping experiment is given), and the resulting model can be used at testing to make inferences about the environment, including estimating the most likely action, based on evidence from other pieces of information.

Several neuroscience and psychology studies build upon the theory of mirror neurons which we brought up in the Introduction. These studies indicate 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 personal experience of the actions being observed in others [16, 17]. This has also been exploited under the deep learning paradigm [18], by using a Multiple Timescales Recurrent Neural Network (MTRNN) to have an artificial simulated agent infer human intention from joint information about object affordances and human actions. One difference between this line of research and ours is that we use real, noisy data acquired from robots and sensors to test our models, rather than virtual simulations.

3. Proposed Approach

In this paper, we combine (1) the robot affordance model of [10], which associates verbal descriptions to the physical interactions of an agent with the environment, with (2) the gesture recognition system of [4], which infers the type of action from human user movements. We consider three *manipulative gestures* corresponding to physical actions performed by agent(s) onto objects on a table (see Fig. 1): grasp, tap, and touch. We reason on the effects of these actions onto the objects of the world, and on the co-occurring verbal description of the experiments. In the complete framework, we will use

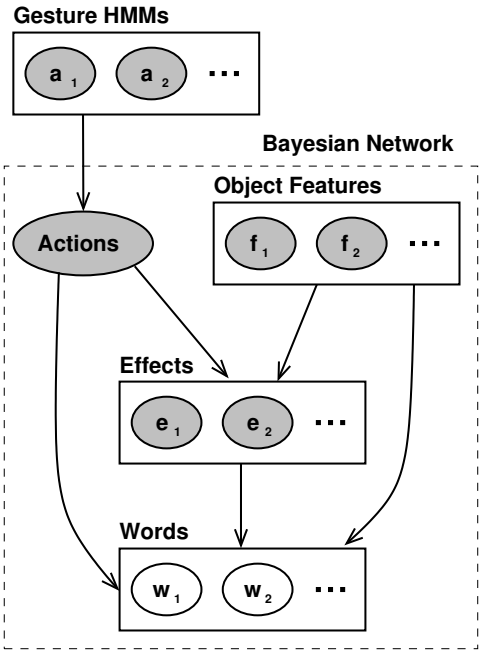


Figure 2: *Abstract representation of the probabilistic dependencies in the model. Shaded nodes are observable or measurable in the present study, and edges indicate Bayesian dependency.*

Bayesian Networks (BNs), which are a probabilistic model that represents random variables and conditional dependencies on a graph, such as in Fig. 2. One of the advantages of using BNs is that their expressive power allows the marginalization over any set of variables given any other set of variables.

Our main contribution is that of extending [10] by relaxing the assumption that the action is known during the learning phase. This assumption is acceptable when the robot learns through self-exploration and interaction with the environment, but must be relaxed if the robot needs to generalize the acquired knowledge through the observation of another (human) agent. We estimate the action performed by a human user during a human–robot collaborative task, by employing statistical inference methods and Hidden Markov Models (HMMs). This provides two advantages. First, we can infer the executed action during training. Secondly, at testing time we can merge the action information obtained from gesture recognition with the information about affordances.

3.1. Bayesian Network for Affordance–Words Modeling

Following the method adopted in [10], we use a Bayesian probabilistic framework to allow a robot to ground the basic world behavior and verbal descriptions associated to it. The world behavior is defined by random variables describing: the actions A, defined over the set $\mathcal{A} = \{a_i\}$, object properties F, over $\mathcal{F} = \{f_i\}$, and effects E, over $\mathcal{E} = \{e_i\}$. We denote $X = \{A, F, E\}$ the state of the world as experienced by the robot. The verbal descriptions are denoted by the set of words $W = \{w_i\}$. Consequently, the relationships between words and concepts are expressed by the joint probability distribution $p(X, W)$ of actions, object features, effects, and words in the spoken utterance. The symbolic variables and their discrete values are listed in Table 1. In addition to the symbolic variables, the model also includes word variables, describing

Table 1: *The symbolic variables of the Bayesian Network which we use in this work (a subset of the ones from [10]), with the corresponding discrete values obtained from clustering during previous robot exploration of the environment.*

name	description	values
Action	action	grasp, tap, touch
Shape	object shape	sphere, box
Size	object size	small, medium, big
ObjVel	object velocity	slow, medium, fast

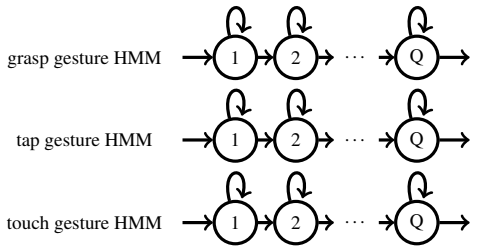


Figure 3: *Structure of the HMMs used for human gesture recognition, adapted from [4]. In this work, we consider three independent, multiple-state HMMs, each of them trained to recognize one of the considered manipulation gestures.*

the probability of each word co-occurring in the verbal description associated to a robot experiment in the environment.

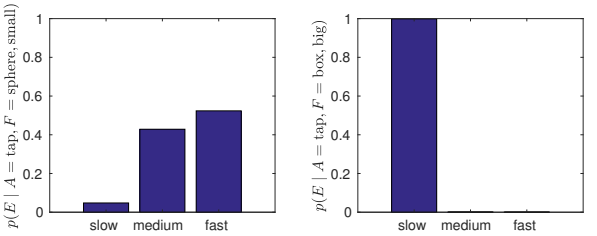
This joint probability distribution, that is illustrated by the part of Fig. 2 enclosed in the dashed box, is estimated by the robot in an ego-centric way through interaction with the environment, as in [10]. As a consequence, during learning, the robot knows what action it is performing with certainty, and the variable A assumes a deterministic value. This assumption is relaxed in the present study, by extending the model to the observation of external (human) agents as explained below.

3.2. Hidden Markov Models for Gesture Recognition

As for the gesture recognition HMMs, we use the models that we previously trained in [4] for spotting the manipulation-related gestures under consideration. Our input features are the 3D coordinates of the tracked human hand: the coordinates are obtained with a commodity depth sensor, then transformed to be centered on the person torso (to be invariant to the distance of the user from the sensor) and normalized to account for variability in amplitude (to be invariant to wide/emphatic vs narrow/subtle executions of the same gesture class).

The gesture recognition models are represented in Fig. 3, and correspond to the Gesture HMMs block in Fig. 2. The HMM for one gesture is defined by a set of (hidden) discrete states $\mathcal{S} = \{s_1, \ldots, s_Q\}$ which model the temporal phases comprising the dynamic execution of the gesture, and by a set of parameters $\lambda = \{A, B, \Pi\}$, where $A = \{a_{ij}\}$ is the transition probability matrix, a_{ij} is the transition probability from state s_i at time t to state s_j at time $t+1$, $B = \{f_i\}$ is the set of Q observation probability functions (one per state i) with continuous mixtures of Gaussian values, and Π is the initial probability distribution for the states.

At recognition (testing) time, we obtain likelihood scores of a new gesture being classified with the common Forward–Backward inference algorithm. In Sec. 3.3, we discuss different ways in which the output information of the gesture recognizer can be combined with the Bayesian Network of words and affordances.



(a) *Prediction of the movement effect on a small sphere.* (b) *Prediction of the movement effect on a big box.*

Figure 4: *Object velocity predictions, given prior information (from Gesture HMMs) that the human user performs a tapping action.*

3.3. Combining the BN with Gesture HMMs

In this study we wish to generalize the model of [10] by observing external (human) agents, as shown in Fig. 1. For this reason, the full model is now extended with a perception module capable of inferring the action of the agent from visual inputs. This corresponds to the Gesture HMMs block in Fig. 2. The Affordance–Words Bayesian Network (BN) model and the Gestures HMMs may be combined in different ways [19]:

1. the Gesture HMMs may provide a hard decision on the action performed by the human (i. e., considering only the top result) to the BN,
2. the Gesture HMMs may provide a posterior distribution (i. e., soft decision) to the BN,
3. if the task is to infer the action, the posterior from the Gesture HMMs and the one from the BN may be combined as follows, assuming that they provide independent information:

$$p(A) = p_{\text{HMM}}(A)\, p_{\text{BN}}(A).$$

In the experimental section, we will show that what the robot has learned subjectively or alone (by self-exploration, knowing the action identity as a prior [10]), can subsequently be used when observing a new agent (human), provided that the actions can be estimated with Gesture HMMs as in [4].

4. Experimental Results

We present preliminary examples of two types of results: predictions over the effects of actions onto environment objects, and predictions over the associated word descriptions in the presence or absence of an action prior. In this section, we assume that the Gesture HMMs provide the discrete value of the recognized action performed by a human agent (i. e., we enforce a hard decision over the observed action, referring to the possible combination strategies listed in Sec. 3.3).

4.1. Effect Prediction

From our combined model of words, affordances and observed actions, we report the inferred posterior value of the Object Velocity effect, given prior information about the action (provided

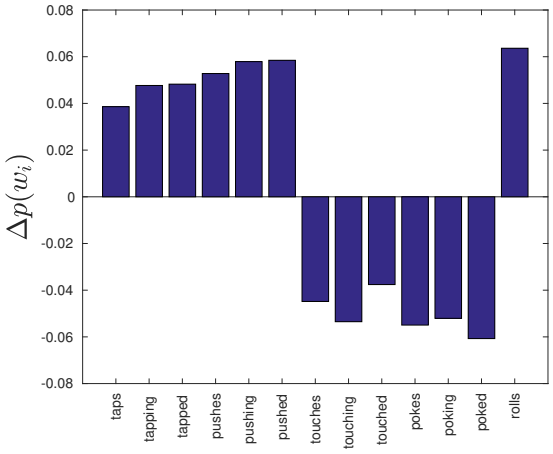


Figure 5: *Variation of word occurrence probabilities:* $\Delta p(w_i) = p(w_i \mid F, E, A = tap) - p(w_i \mid F, E)$*, where* $F =$ {*Size=big, Shape=sphere*}*,* $E = \{$*ObjVel=fast*$\}$*. This variation corresponds to the difference of word probability when we add the tap action evidence (obtained from the Gesture HMMs) to the initial evidence about object features and effects. We have omitted words for which no significant variation was observed.*

by the Gesture HMMs) and also about object features (Shape and Size). Fig. 4 shows the computed predictions in two cases. Fig. 4a shows the anticipated object velocity when the human user performs the tapping action onto a small spherical object, whereas Fig. 4b displays it when the target object is a big box. Indeed, given the same observed action prior (lateral tap on the object), the expected movement is very different depending on the physical properties of the target object.

4.2. Prediction of Words

In this experiment, we compare the associated *verbal description* obtained by the Bayesian Network in the absence of an action prior, with the ones obtained in the presence of one. In particular, we compare the *probability of word occurrence* in the following two situations:

1. when the robot prior knowledge (evidence in the BN) includes information about object features and effects only: *Size=big, Shape=sphere, ObjVel=fast*;
2. when the robot prior knowledge includes, in addition to the above, evidence about the action as observed from the Gestures HMMs: *Action=tap*.

Fig. 5 shows the variation in word occurrence probabilities between the two cases, where we have omitted words for which no significant variation was observed in this case. We can interpret the difference in the predictions as follows:

- as expected, the probabilities of words related to tapping and pushing increase when a tapping action evidence from the Gestures HMMs is introduced; conversely, the probabilities of other action words (touching and poking) decreases;
- interestingly, the probability of the word *rolling* (which is an effect of an action onto an object) also increases when the tapping action evidence is entered. Even though the initial evidence of case 1 already included some effect information (the velocity of the object), it is only now, when the robot perceives that the physical action was a tap, that the event rolling is associated.

5. Conclusions and Future Work

Within the scope of cognitive robots that operate in unstructured environments, we have discussed a model that combines word affordance learning with body gesture recognition. We have proposed such an approach, based on the intuition that a robot can generalize its previously-acquired knowledge of the world (objects, actions, effects, verbal descriptions) to the cases when it observes a human agent performing familiar actions in a shared human–robot environment. We have shown promising preliminary results that indicate that a robot's ability to predict the future can benefit from incorporate the knowledge of a partner's action, facilitating scene interpretation and, as a result, teamwork.

In terms of future work, there are several avenues to explore. The main ones are (i) the implementation of a fully probabilistic fusion between the affordance and the gesture components (e. g., the soft decision discussed in Sec. 3.3); (ii) to run quantitative tests on larger corpora of human–robot data; (iii) to explicitly address the correspondence problem of actions between two agents operating on the same world objects (e. g., a pulling action from the perspective of the human corresponds to a pushing action from the perspective of the robot, generating specular effects).

6. Acknowledgements

This research was partly supported by the CHIST-ERA project IGLU and by the FCT project UID/EEA/50009/2013. We thank Konstantinos Theofilis for his software and help permitting the acquisition of human hand coordinates in human–robot interaction scenarios with the iCub robot.

7. References

[1] B. Siciliano and O. Khatib, *Springer Handbook of Robotics*, 2nd ed. Springer, 2016.

[2] C. Matuszek, L. Bo, L. Zettlemoyer, and D. Fox, "Learning from Unscripted Deictic Gesture and Language for Human–Robot Interactions," in *AAAI Conference on Artificial Intelligence*, 2014, pp. 2556–2563.

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

[4] G. Saponaro, G. Salvi, and A. Bernardino, "Robot Anticipation of Human Intentions through Continuous Gesture Recognition," in *International Conference on Collaboration Technologies and Systems*, ser. International Workshop on Collaborative Robots and Human–Robot Interaction, 2013, pp. 218–225.

[5] G. Rizzolatti, L. Fogassi, and V. Gallese, "Neurophysiological mechanisms underlying the understanding and imitation of action," *Nature Reviews Neuroscience*, vol. 2, pp. 661–670, 2001.

[6] J. J. Gibson, *The Ecological Approach to Visual Perception: Classic Edition*. Psychology Press, 2014, originally published in 1979 by Houghton Mifflin Harcourt.

[7] L. Montesano, M. Lopes, A. Bernardino, and J. Santos-Victor, "Learning Object Affordances: From Sensory–Motor Maps to Imitation," *IEEE Transactions on Robotics*, vol. 24, no. 1, pp. 15–26, 2008.

[8] L. Jamone, E. Ugur, A. Cangelosi, L. Fadiga, A. Bernardino, J. Piater, and J. Santos-Victor, "Affordances in psychology, neuroscience and robotics: a survey," *IEEE Transactions on Cognitive and Developmental Systems*, 2016.

[9] N. Krüger, C. Geib, J. Piater, R. Petrick, M. Steedman, F. Wörgötter, A. Ude, T. Asfour, D. Kraft, D. Omrčen, A. Agostini, and R. Dillmann, "Object–Action Complexes: Grounded Abstractions of Sensory–Motor Processes," *Robotics and Autonomous Systems*, vol. 59, no. 10, 2011.

[10] G. Salvi, L. Montesano, A. Bernardino, and J. Santos-Victor, "Language Bootstrapping: Learning Word Meanings From Perception–Action Association," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42, no. 3, pp. 660–671, 2012.

[11] A. F. Morse and A. Cangelosi, "Why Are There Developmental Stages in Language Learning? A Developmental Robotics Model of Language Development," *Cognitive Science*, vol. 41, pp. 32–51, 2016.

[12] F. Stramandinoli, V. Tikhanoff, U. Pattacini, and F. Nori, "Grounding Speech Utterances in Robotics Affordances: An Embodied Statistical Language Model," in *IEEE International Conference on Developmental and Learning and on Epigenetic Robotics*, 2016, pp. 79–86.

[13] A. Gonçalves, G. Saponaro, L. Jamone, and A. Bernardino, "Learning Visual Affordances of Objects and Tools through Autonomous Robot Exploration," in *IEEE International Conference on Autonomous Robot Systems and Competitions*, 2014.

[14] A. Gonçalves, J. Abrantes, G. Saponaro, L. Jamone, and A. Bernardino, "Learning Intermediate Object Affordances: Towards the Development of a Tool Concept," in *IEEE International Conference on Developmental and Learning and on Epigenetic Robotics*, 2014.

[15] 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 *IEEE International Conference on Robotics and Automation*, 2016.

[16] S. M. Aglioti, P. Cesari, M. Romani, and C. Urgesi, "Action anticipation and motor resonance in elite basketball players," *Nature Neuroscience*, vol. 11, no. 9, pp. 1109–1116, 2008.

[17] G. Knoblich and R. Flach, "Predicting the Effects of Actions: Interactions of Perception and Action," *Psychological Science*, vol. 12, no. 6, pp. 467–472, 2001.

[18] S. Kim, Z. Yu, and M. Lee, "Understanding human intention by connecting perception and action learning in artificial agents," *Neural Networks*, vol. 92, pp. 29–38, 2017.

[19] R. Pan, Y. Peng, and Z. Ding, "Belief Update in Bayesian Networks Using Uncertain Evidence," in *IEEE International Conference on Tools with Artificial Intelligence*, 2006, pp. 441–444.