Predicting Object Affordances within a Continuous Dialogue State Update Process for Human-Robot-Interaction

Julian Hough and Lorenzo Jamone

***Abstract*—We present a flexible, general model for a robot predicting affordances of potentially unfamiliar objects in a human-robot dialogue system with incremental speech understanding capabilities. We show how predicting affordances needs to be integrated within a general, continuous dialogue state update process in order to take advantage of dialogue-level information in tandem with other perceptual information. We outline how this is done using probabilistic type theory whereby incoming sensory data is converted to a world belief record in real time, and then derived beliefs such as intention attribution to a user, or the prediction of affordances of objects, are made as probabilistic record type judgements of that record.**

I. Introduction

Collaborative robots with speech understanding capabilities, which are designed for real-world human-robot interaction (HRI) need to combine numerous sources of information in their current world belief state to carry out joint tasks effectively and fluidly with users. These information sources include probabilistic visual and physical property judgements of objects and information about the interaction commonly encoded in a *dialogue state*. A uniform approach not only requires the use of complex visual information and semantic parsing, but needs to permit fluid interaction with a collaborative robot to help a user complete a manual task. This requires an incrementally and dynamically evolving world belief which encodes the robot's own action state as well as its estimation of the user's intentions in real time.

In this paper we address this challenge by formulating a simple interaction state for a robot using concepts from Type Theory with Records (TTR) [1]. We characterize the robot's world belief as a constantly updating *record*, and use *record type classifiers* of different kinds which operate on the state record to make type judgements on the world belief. Once a judgement is made and used (committed), this can be added to the world belief for further classification and update. For the classification we use a combination of lattice theory and probabilistic TTR [2]. Inspired by the recent work using TTR for perceptual classification [4, 11] and the simple Words-As-Classifiers (WAC) model [8] to reference resolution of objects in real-world scenes, here we propose a general Types-As-Classifiers (TAC) approach.

II. Types-As-Classifiers for affordance prediction in human-robot interaction

Typical raw perceptual information for a collaborative pick-and-place robot may be as in the left of in Fig. 1. This shows a camera feed, and computer vision based segmentation and tracking of objects as described in [10], and perceptual classifiers, such as that for 'yellow', which classify the degree to which an object has that perceptual property. The current words recognized by the robot's speech recognizer (ASR) are also added to the state as they arrive. The robot action state diagram shows the robot tracking its own current task and action state through a Hierarchical State Machine (HSM).

To encode the current state with all the incoming sensory information, our system uses TTR *record types*, and the inhabitants of record types, *records*, as our primary formal apparatus – see [1] for details. We characterize the state as a *world belief record*- for an our in-robot system it will be of the format of record wb on the right of Fig. 1.[1]

The driving incremental interpretation process of the system is a probabilistic classification of the current world belief record wb as being of a given situation record type i within a set of possible record types I, conditioned by current evidence record type e– i.e. $p(wb : i|wb : e)$. See an example record type judgement i on the right of Fig. 1. Different perceptual classifiers operate on wb to yield probability judgements that wb is of a given type, such as probabilistic parsing of incoming words and user intention recognition. Once a type judgement is commited that wb is of a given record type i, this is added to wb, and then further judgements of its type can be made and added to it. We are not committed to a specific classification ordering or algorithm here, and leave investigation into this for future work. However, we are committed to the distribution over possible record type judgements being stored in a record type lattice– see [7].

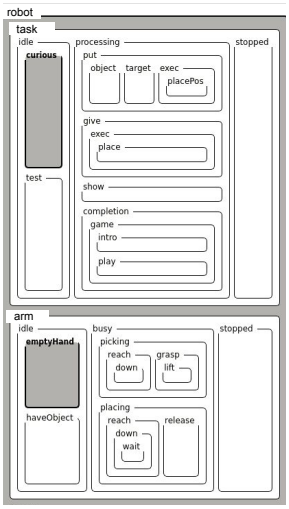
While many different perceptual type classifications can operate on wb to update it, here we focus on the perception of object *affordances* [5], i.e. the possible actions associated to the objects (e.g. *graspable*), which are crucial for the robot to be able to manipulate them. Recently, probabilistic computational models of affordance perception have been proposed using Bayesian Networks [6] and variational auto-encoders [3]- these can be used to obtain the probability of an object having different affordances from visual and linguistic features. In our model, affordance prediction is part of the probabilistic type judgement of wb, such that the probabilities of each object having each affordance property

[1]This is an example record where many of the labels and values are just represented by '...' to indicate at least one such field would be present in the full representation.

SCENE:

OBJECTS (segmentation + visual classifiers):
object_0:
yellow = 0.69
blue = 0.38
round = 0.66
..
object_1:
yellow = 0.10
blue = 0.86
round = 0.34
...
USER SPEECH:
'put the left green apple in the basket'

ROBOT ACTION STATE:



WORLD BELIEF RECORD:

$$wb = \begin{bmatrix} objects = \begin{bmatrix} obj\_0 = [\ ... = ...\ ] \\ obj\_1 = [\ ... = ...\ ] \\ ... = ... \\ obj\_N = [\ ... = ...\ ] \end{bmatrix} \\ robot = \begin{bmatrix} arm = [\ ... = ...\ ] \\ task = [\ ... = ...\ ] \\ intention = [\ ... = ...\ ] \end{bmatrix} \\ human = \begin{bmatrix} c{-}utt = [\ parse = ...\ words = ...\ ] \\ status = ... \\ intention = [\ ... = ...\ ] \end{bmatrix} \end{bmatrix}$$

PROPERTY AND AFFORDANCE JUDGEMENT:

$$i = \begin{bmatrix} objects : \begin{bmatrix} obj\_1 : \begin{bmatrix} x : e \\ green : green(x) \\ round : round(x) \\ graspable : graspable(x) \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Fig. 1. LEFT: A typical state according to the robot. Objects are segmented and properties can be obtained for each object. The robot's internal action state is controlled by a Hierarchical State Machine (HSM). RIGHT: The format of the world belief record which represents the state, and below, an example record type which our probabilistic classifier assigns a probability value to given the current record.

are part of the available type judgements. For example, the field $[\, graspable : graspable(x) \,]$ in the example record type judgement i in Fig. 1 is the judgement that obj_1 is graspable, based on current information elsewhere in wb. Notice we combine this judgement with colour and shape classification, as our approach does not commit to separating judgements arbitrarily, allowing for relevant type judgements to be learned. Many other combinations of fields are possible for record type judgements on wb, and these can be calculated through the probabilistic type theory equations in [7]. Our general approach is to model how affordance prediction can best integrate with natural language processing decisions as in [9], but here integrated into a continuous dialogue state update.

References

[1] Robin Cooper. Records and record types in semantic theory. *Journal of Logic and Computation*, 15(2), 2005.

[2] Robin Cooper, Simon Dobnik, Shalom Lappin, and Staffan Larsson. A probabilistic rich type theory for semantic interpretation. In *Proceedings of the EACL Workshop on Type Theory and Natural Language Semantics (TTNLS)*, Gothenburg, Sweden, 2014. ACL.

[3] Atabak Dehban, Lorenzo Jamone, Adam R Kampff, and José Santos-Victor. Denoising auto-encoders for learning of objects and tools affordances in continuous space. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, pages 4866–4871. IEEE, 2016.

[4] Simon Dobnik, Robin Cooper, and Staffan Larsson. Modelling language, action, and perception in type theory with records. In *International Workshop on Constraint Solving and Language Processing*, pages 70–91. Springer, 2012.

[5] James J Gibson. The theory of affordances. *The people, place, and space reader*, pages 56–60, 1979.

[6] Afonso Gonçalves, João Abrantes, Giovanni Saponaro, Lorenzo Jamone, and Alexandre Bernardino. Learning intermediate object affordances: Towards the development of a tool concept. In *Joint IEEE International Conferences on Development and Learning and Epigenetic Robotics (ICDL-Epirob), 2014*, pages 482–488. IEEE, 2014.

[7] Julian Hough and Matthew Purver. Probabilistic record type lattices for incremental reference processing. In *Modern perspectives in type-theoretical semantics*, pages 189–222. Springer, 2017.

[8] Casey Kennington and David Schlangen. Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. Proceedings of the Conference for the Association for Computational Linguistics (ACL). ACL, 2015.

[9] 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)*, 42(3):660–671, 2012.

[10] André Ückermann, Christof Elbrechter, Robert Haschke, and Helge Ritter. Hierarchical Scene Segmentation and Classification. In *Robots in Clutter Workshop at IROS 2014*, 2014.

[11] Yanchao Yu, Arash Eshghi, and Oliver Lemon. Training an adaptive dialogue policy for interactive learning of visually grounded word meanings. In *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 339, 2016.