# A system for interactive learning in dialogue with a tutor


Danijel Skočaj, Matej Kristan, Alen Vrečko, Marko Mahnič — University of Ljubljana, Slovenia

Miroslav Janíček, Geert-Jan M. Kruijff — DFKI, Saarbrücken, Germany

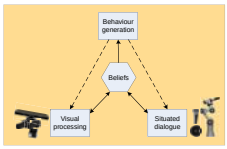
Michael Zillich, Kai Zhou — TU Vienna, Austria

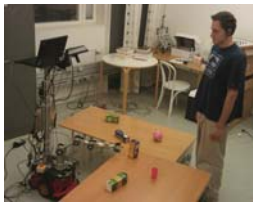
Marc Hanheide, Nick Hawes — Uni. of Birmingham, UK

Thomas Keller — Albert-Ludwigs-Uni. Freiburg, Germany


## 1. Introduction

**Interactive continuous learning** is an important characteristic of a cognitive agent that is supposed to operate and evolve in an everchanging environment. We present **representations** and **mechanisms** that are necessary for continuous learning of visual concepts in **dialogue with a tutor.** We present an approach for **modelling beliefs** and we show how these beliefs are created by processing **visual and linguistic information**. Based on the detected **knowledge gaps** represented in the beliefs, the **motivation and planning** mechanism implements four types of interaction for **learning**. These principles have been implemented in an **integrated system**.



## 2. Visual processing

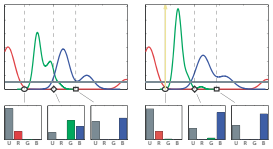
Visual processing serves to provide the **object hypotheses** together with **visual properties** about which the system will subsequently learn. Given that the system learns from a variety of as yet unknown objects, we implemented a generic segmentation scheme, exploiting the fact that objects are presented on planar supporting surfaces [1]. The vision subsystem is an **active observer** using a **wide field of view** Kinect sensor and a pair of **narrow field of view** stereo cameras for foveated vision, both mounted on a **pan-tilt unit** (PTU).

The objects are detected based on the **plane-pop-out** approach using the Kinect 3D point cloud. Then, every object is attended by moving the PTU accordingly, and **segmented** in the higher resolution 2D image using the graph-cut algorithm initialized by 2D and 3D data. The features are then extracted from the segmented image regions and corresponding 3D data, which are then used for **recognition** and **learning** of objects and their colors and shapes.

## 3. Learning visual concepts

The visual concepts are represented as generative models that take the form of probability density functions over the feature space. They are based on multivariate **online discriminative Kernel Density Estimator** (odKDE) [2] and are constructed in an online fashion from new observations by adapting from the positive examples (**learning**) as well as negative examples (**unlearning**) and by taking into account the probability that a concept that has not been observed before has been encountered by maintaining the representation of the **unknown model**.

**Detection of incompleteness in knowledge**:
AP of $M_0$ is the best -> knowledge gap
AP of the best $M_i$ is low -> uncertainty



## 4. Modelling beliefs and intentions

**Beliefs** express factual information about the state of the world. In our approach, they are relational structures that account for the inherent uncertainty using multivariate probability distributions over properties and their values. They are situated, anchored to a given situation, and mutually interlinked. We model three degrees of belief attribution, which we call the epistemic status: private, attributed and shared. **Private beliefs** are internal to the robot, and are usually the result of perception or deliberation. **Attributed beliefs** are beliefs that other agents expressed by communicative means. Finally, **shared beliefs** form the common ground established in the interaction.

**Intentions**, on the other hand, are closely related to rational aspects of interaction. Behind every (intentional) action, there is an underlying intention. For instance, when asking a question, the intention is to elicit an answer, i.e. to get to a state in which the question is answered. We use intentions as a unified representation for actions of both the robot and the human.

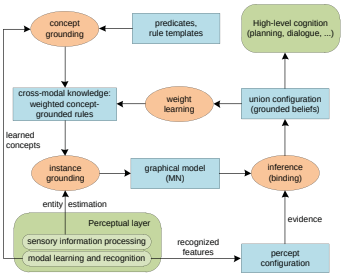
## 5. Situated dialogue

Situated dialogue understanding and production is treated as an **abductive problem**. Language **understanding** is treated as inference to the most appropriate intention and beliefs behind a communicative act, whereas **production** is inference to the most appropriate realization of the robot's (communicative) intention and beliefs.

Given a goal, the abductive reasoner builds up and continually refines a set of partial defeasible explanations of the input, conditioned on the verification of the knowledge gaps they contain. This verification is done by executing test actions, thereby going beyond the initial context [3].

## 6. Binding and reference resolution

**Binding** - the ability to combine two or more modal representations of the same entity into a single shared representation is vital for every cognitive system operating in a complex environment. **Reference resolution** is a process akin to binding that relates information attributed to another agent to the robot's own perceptions. We developed a general probabilistic binding method based on **Markov Logic Networks** and applied it to the problem of reference resolution in our cognitive system [4].




## 7. Behaviour generation

The **motivation management** [5] monitors the beliefs and based on them creates **goals** and selects which of them to pass on to planning. The **planner** [6] then builds a plan to satisfy a given goal, which is subsequently executed. In this way a system behaviour is generated and controlled.

The system switches between **different behaviours:**

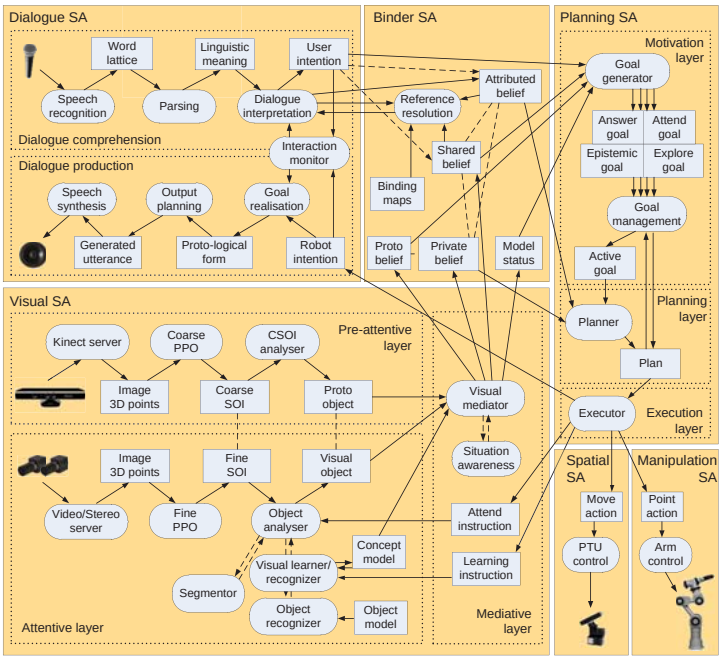
**Answer questions:**
Answer the question verbally.
Point at an object.
**Learn object properties:**
Invoke different learning mechanisms.
**Look around:**
Look around the scene and try to recognize all objects.

Implemented **Learning mechanims**:

**Situated tutor driven learning**
The human drives the learning.
H: "The box is red."
**Situated tutor assisted learning**
The robot takes the initiative.
R: "Is this yellow?"
**Non-situated tutor assisted learning**
Introspection - model analysis.
R: "Could you show me something red?"
**Autonomous learning**
Robot automatically updates the models.

## 8. The system

The system architecture is based on **CAS** (CoSy Architecture Schema) [7]. The schema is essentially a **distributed working memory model**, where representations are linked within and across the working memories, and are updated asynchronously and in parallel. Using this architecture, a complex, distributed, asynchronous, and heterogenuous system has been built [8].




## References


[1] K. Zhou et.al. (2011). Visual Information Abstraction For Interactive Robot Learning. In proceedings of ICAR 2011, pages 328-334, Tallin, Estonia.
[2] M. Kristan and A. Leonardis (2010). Online Discriminative Kernel Density Estimation. In Proceedings of the ICPR 2010, pages 581-584, Istanbul, Turkey.
[3] M. Janíček (2011). Abductive Reasoning for Continual Dialogue Understanding. In Proceedings of the ESSLLI 2011, Ljubljana, Slovenia.
[4] A. Vrečko, A. Leonardis, and D. Skočaj (2012). Modeling Binding and Cross-modal Learning in Markov Logic Networks. Neurocomputing.
[5] M. Hanheide et.al. (2010). A Framework for Goal Generation and Management. In Proc. of the AAAI Workshop on Goal-Directed Autonomy, Atlanta, Georgia.
[6] M. Brenner and B. Nebel (2009). Continual planning and acting in dynamic multiagent environments. JAAMAS, 19(3):297-331.
[7] N. Hawes and J. Wyatt (2010). Engineering intelligent information processing systems with CAST. Advanced Engineering Infomatics, 24(1):27–39, 2010.
[8] D. Skočaj et. al. (2011). A system for interactive learning in dialogue with a tutor. In IROS 2011, pages 3387 – 3394, San Francisco, CA, USA.


Video at http://cogx.eu/results/george

## Acknowledgment


This work was supproted by the EC FP7 IST project CogX-215181