A system approach to interactive learning of visual concepts

Danijel Skočaj[1] **Matej Kristan**[1] **Aleš Leonardis**[1] **Marko Mahnič**[1]
Alen Vrečko[1] **Miroslav Janíček**[2] **Geert-Jan M. Kruijff**[2] **Pierre Lison**[2]
Michael Zillich[3] **Charles Gretton**[4] **Marc Hanheide**[4]
Moritz Göbelbecker[5]

[1]University of Ljubljana, Slovenia [2]DFKI, Saarbrücken, Germany
[3]Vienna University of Technology, Austria [4]University of Birmingham, UK
[5]Albert-Ludwigs-Universität Freiburg, Germany

Abstract

In this work we present a system and underlying mechanisms for continuous learning of visual concepts in dialogue with a human.

1. Introduction

Cognitive systems are often characterised by their ability to learn, communicate and act autonomously. In combining these competencies we envision a system that incrementally learns about the scene by being engaged in mixed initiative dialogues with a human tutor. In this paper we outline how our robot George, depicted in Fig. 1, learns and refines visual conceptual models, either by attending to information deliberatively provided by a human tutor (*tutor-driven learning*: e.g., H: This is a red box.) or by taking initiative itself, asking the tutor for specific information (*tutor-assisted learning*: e.g., G: Is the elongated object yellow?). Our approach unifies these cases into an integrated approach comprising incremental visual learning, selection of learning goals, continual planning to select actions for optimal learning behaviour, and a dialogue subsystem. George is one system in a family of memory-oriented integrated systems that aim to understand where their own knowledge is incomplete and that take actions to extend their knowledge subsequently. Our objective is to demonstrate a cognitive system that can efficiently acquire conceptual models in an interactive learning process that is not overly taxing with respect to tutor supervision and is performed in an intuitive, user-friendly way.

2. The system

The implementation of the robot is based on CAS (Hawes and Wyatt, 2010), the CoSy Architecture Schema. The schema is essentially a distributed working memory model composed of several subarchitectures (SAs) implementing different functionalities. George is composed of four such SAs, as depicted in Fig. 1 (here, the components are depicted as rounded boxes and exchanged data structures as rectangles, with arrows indicating a conceptual information flow).

The **Visual SA** processes the scene as a whole using stereo pairs of images and identifies spaces of interest, which are further analysed; the potential objects are segmented and are then subjected to feature extraction. The extracted features are then used for **learning and recognition** of objects and qualitative visual attributes, like colour and shape. The learning is based on an online learning method that enables updating from positive examples (*learning*) as well as from negative examples (*unlearning*) (Kristan et al., 2010). Our approach also does not assume the closed world assumption; at every step the system also takes into account the probability that it has encountered a concept that has not been observed before.

The recognised visual properties are then forwarded to the **Binder SA**, which serves as a central hub for gathering information from different modalities about entities currently perceived in the environment. Since the information was extracted by the robot itself, we call the resulting information structure a *private belief*.

Beliefs can also be created by the **Dialogue SA**. It analyses an incoming audio signal, parses the created word lattice and chooses the contextually most appropriate meaning representation for the utterance (Lison and Kruijff, 2009). Then it establishes which meaningful parts might be referring to objects in the visual context. The actual reference resolution then takes place when we perform dialogue interpretation taking into account the information stored in the robot beliefs. In this process, we use weighted abductive inference to establish the intention behind the utterance. As a result of this process, an *attributed belief* containing the information asserted by the human is constructed from the meaning of

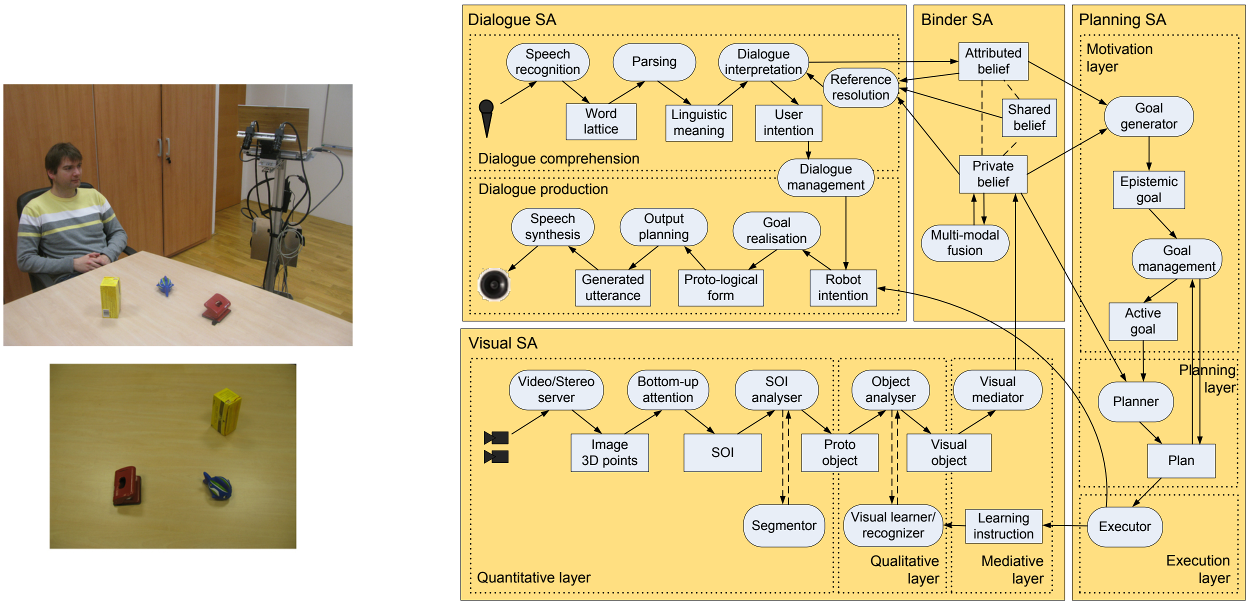

Figure 1: Left: Scenario setup and observed scene. Right: Schematic system architecture.

the utterance.

The beliefs, being high-level symbolic representations, provide a unified model of the environment and the attributed information, which can be efficiently used for planning. In the **Planning SA**, the *motivation layer* (Hanheide and et.al., 2010) monitors beliefs to generate goals. In the tutor-driven case attributed beliefs are taken as learning opportunities eventually leading to epistemic goals that require that this new information is used to update the visual models. To implement robot-initiated tutor assisted learning the Planning SA also continuously accounts for the goal to maximise the system's knowledge in terms of reducing the uncertainty in the visual models. In order to only take action if there is a significant learning opportunity in terms of reward, goal management employs a threshold to decide whether a plan should be executed. The goal management continuously manages goals according to their priority, eventually interrupting execution if a higher priority goal shows up. We assign human-initiated goals a higher priority, enabling the system to immediately respond to human input.

Plan *execution* proceeds according to the continual planning paradigm (Brenner and Nebel, 2009) monitoring the system's belief to trigger replanning if required. In tutor driven learning, actions scheduled for execution typically include sending a learning instruction to the Visual SA, which triggers the update of the visual representations. In tutor assisted learning the execution usually involves sending a clarification request to the Dialogue SA, which is then subsequently synthesised, typically as a polar or open question about a certain object property, and the tutor's answer is then used to update the models.

3. Conclusion

In this work we briefly presented the integrated system and underlying mechanisms for continuous learning of visual concepts in dialogue with a human tutor. Building on this system, our final goal is to produce an autonomous robot that will be able to efficiently learn and act by capturing and processing cross-modal information in interaction with the environment and other cognitive agents.

Acknowledgment

The work was supported by the EC FP7 IST project CogX-215181.

References

Brenner, M. and Nebel, B. (2009). Continual planning and acting in dynamic multiagent environments. *JAAMAS*, 19(3):297–331.

Hanheide, M., et.al. (2010). A framework for goal generation and management. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Hawes, N. and Wyatt, J. (2010). Engineering intelligent information-processing systems with CAST. *Advanced Engineering Infomatics*, 24(1):27–39.

Kristan, M., et.al. (2010). Online kernel density estimation for interactive learning. *Image and Vision Computing*, 28(7):1106–1116.

Lison, P. and Kruijff, G. (2009). Efficient parsing of spoken inputs for human-robot interaction. In *Proceedings of the RO-MAN 09*, Toyama, Japan.