A basic cognitive system for interactive learning of simple visual concepts

Danijel Skočaj, Matej Kristan, Aleš Leonardis, Alen Vrečko
University of Ljubljana, Slovenia

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

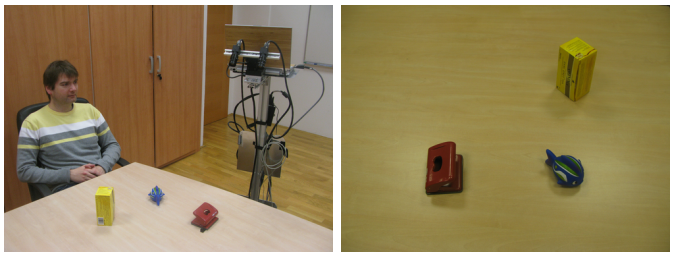
Michael Zillich
Vienna University of Technology, Austria

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

I. Introduction

Two common and important characteristics of cognitive systems are the ability to learn and the ability to communicate. By combining both competencies such a system could also be capable of interactive learning, i.e., learning in dialogue with a human, which should significantly facilitate the incremental learning processes. In this work we briefly describe the representations and mechanisms that enable such interactive learning and present a system that was designed to acquire visual concepts through interaction with a human [1].

Fig. 1 depicts our robot George engaged in a dialogue with a human tutor[1]. In this scenario, the main goal is to teach the robot about object properties (colours and two basic shapes) in an interactive way. The tutor can teach the robot about object properties (e.g., 'H: This is a red thing.'), or the robot can try to learn autonomously or ask the tutor for help when necessary (e.g., 'G: Is the elongated thing red?'). Our aim is that the learning process is efficient in terms of learning progress, is not overly taxing with respect to tutor supervision and is performed in a natural, user friendly way.



(a) Scenario setup. (b) Observed scene.

Fig. 1. Continuous interactive learning of visual properties.

To enable interactive learning, the system has to be able to, on one hand, perceive the scene and (partially) interpret the visual information and build the corresponding representations of visual objects, and on the other hand, to communicate with the tutor and interpret the tutor's utterances, forming the corresponding representations of the linguistic meaning. The system should then relate these two types of modal representations and on top of them create new, a-modal, representations that enable further communication and allow for incremental updating of visual models, therefore facilitate incremental learning. In the following section we will describe the robot system we have developed, and the underlying representations and mechanisms that implement the premises mentioned above.

[1]The robot can be seen in action in the video accessible at http://cogx.eu/results/george.

II. The system

The implementation of the robot is based on a specific architecture schema, which we call CAS (CoSy Architecture Schema) [2]. 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. The system is therefore composed of several subarchitectures (SAs) implementing different functionalities and communicating through their working memories. The George system is composed of three such subarchitectures: the *Binder SA*, the *Communications SA* and the *Visual SA*, as depicted in Fig. 2.

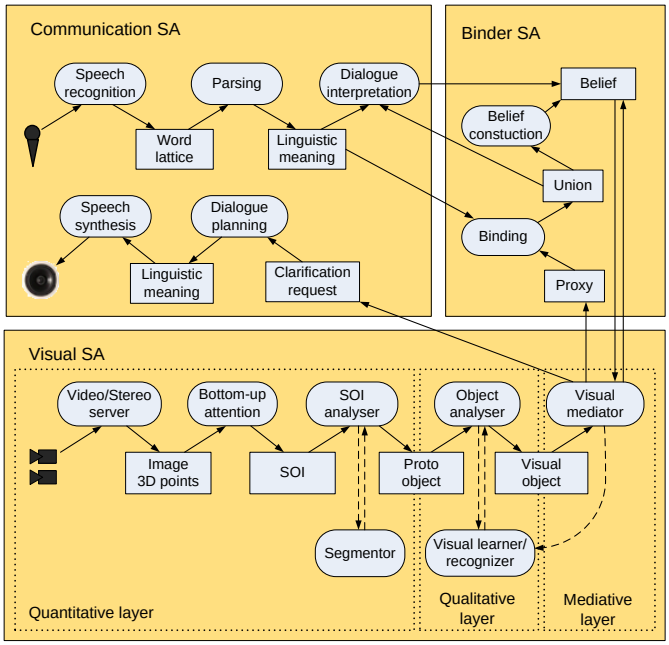


Fig. 2. Architecture of the George system.

The **Visual SA** processes the visual scene as a whole using stereo pairs of images and identifies regions in the scene that might be interesting for further visual processing (3D spaces of interest or SOIs that stick out of the plane). These regions are further analysed; the potential objects (proto objects) are segmented using 3D and colour information and are then subjected to feature extraction. The extracted features are then used for **learning and recognition** of qualitative visual attributes, like colour and shape. These visual concepts are represented as generative models that take the form of probability density functions over the feature space. They are based on *online Kernel Density Estimator (oKDE)* [3], that we have developed, and are constructed in an online fashion from new observations. The oKDE estimates the probability density functions by a mixture of Gaussians, and is able to adapt from the positive examples (*learning*) as well as negative examples (*unlearning*) [4]. Our approach also does not assume the close 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. Therefore, during online operation, a multivariate generative model is continually maintained for each of the visual concepts and for mutually exclusive sets of concepts (e.g., all colours, or all shapes) the optimal feature subspace is continually being determined. This feature subspace is then used to construct a Bayesian classifier for a set of mutually exclusive concepts, which is used for recognition of individual object properties.

The recognized 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. Based on the available information, the binder seeks to fuse the perceptual inputs arising from the various subarchitectures, by checking whether their respective features correlate with each other. The probability of these correlations are encoded in a Bayesian network. We call the resulting (amodal) information structure a *belief*. The task of the binder is to decide which perceptual inputs belong to the same real-world entity, and should therefore be unified into a belief. The decision algorithm uses a technique from probabilistic data fusion, called the *Independent Likelihood Pool* (ILP) [5]. A belief encodes also additional information related to the specific situation and perspective in which the belief was formed, such as spatio-temporal frame, epistemic status and a saliency value. The beliefs, being high-level symbolic representations available for the whole cognitive architecture, provide a unified model of the environment which can be efficiently used when interacting with the human user.

Beliefs can also be altered by **Communication SA**. It analyses an incoming audio signal and parses the created word lattice. From the space of possible linguistic meaning representations for the utterance, the contextually most appropriate one is chosen [6]. We represent this meaning as a *logical form*, an ontologically richly sorted relational structure. Given this structure, the Communication SA 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. In this process, we use weighted abductive inference to establish the intention behind the utterance.

If the intention was to provide to the system a novel information that can be learned (*Tutor driven learning*), a belief attributed to the human is constructed from the meaning of the utterance. This event triggers a learning opportunity in Visual SA, where the corresponding visual concepts are updated. The learning process can also be initiated by the system itself (*Tutor assisted learning*). In the case of missing or ambiguous modal information (interpretation of the current scene), Visual SA can send a clarification request to Communication SA, which formulates a dialogue goal given the information the system needs to know and how that can be related to the current dialogue and belief-context. Dialogue planning turns this goal into a meaning representation that expresses the request in context. This 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.

III. Conclusion

In this work we briefly presented the system and underlying representations and mechanisms for continuous learning of visual concepts in dialogue with a human tutor. We have made several contributions at the level of individual components (modelling beliefs, dialogue processing, incremental learning), as well as at the system level; all the components presented in this work have been integrated into a coherent multimodal distributed asynchronous system. Building on this system, our final goal is to produce an autonomous robot that will be able to efficiently learn and adapt to an everchanging world by capturing and processing cross-modal information in an interaction with the environment and other cognitive agents.

Acknowledgment

The work was supported by the EC FP7 IST project CogX-215181, and partially by the Research program Computer Vision P2-0214 (RS).

References

[1] D. Skočaj, M. Janiček, M. Kristan, G.-J. M. Kruijff, A. Leonardis, P. Lison, A. Vrečko, and M. Zillich, "A basic cognitive system for interactive continuous learning of visual concepts," in *Proceedings of the ICRA 2010 Workshop on Interactive Communication for Autonomous Intelligent Robots (ICAIR) Making robots articulate what they understand, intend, and do.*, Anchorage, AK, USA, May 2010.

[2] N. Hawes and J. Wyatt, "Engineering intelligent information-processing systems with CAST," *Advanced Engineering Infomatics*, vol. 24, no. 1, pp. 27–39, 2010.

[3] M. Kristan and A. Leonardis, "Multivariate online kernel density estimation," in *Computer Vision Winter Workshop*, Nové Hrady, Czech Republic, February 2010, pp. 77–86.

[4] M. Kristan, D. Skočaj, and A. Leonardis, "Online kernel density estimation for interactive learning," *Image and Vision Computing*, vol. 28, no. 7, pp. 1106–1116, July 2010.

[5] E. Punskaya, "Bayesian approaches to multi-sensor data fusion," Master's thesis, Cambridge University Engineering Department, 1999.

[6] P. Lison and G. Kruijff, "Efficient parsing of spoken inputs for human-robot interaction," in *Proceedings of the 18th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 09)*, Toyama, Japan, 2009.