EU FP7 CogX
ICT-215181
May 1 2008 (52months)

DR 7.5:
A curiosity driven self-extending robot system

Danijel Skočaj, Matej Kristan, Alen Vrečko, Marko Mahnič, Miroslav Janíček, Geert-Jan M. Kruijff, Marc Hanheide, Nick Hawes, Thomas Keller, Michael Zillich and Kai Zhou

University of Ljubljana, DFKI Saarbrücken, University of Birmingham, Albert-Ludwigs-Universität Freiburg, TU Vienna
⟨danijel.skocaj@fri.uni-lj.si⟩

Due date of deliverable:	30 June 2012
Actual submission date:	30 May 2012
Lead partner:	UL
Revision:	final
Dissemination level:	PU

In this deliverable we present the curiosity driven self-extending robot system George that is capable of interactive learning of visual concepts in a dialogue with a human tutor. We present representations and mechanisms that facilitate such continuous interactive learning. We present how beliefs about the world are created by processing visual and linguistic information and show how they are used for planning the system behaviour with the aim at satisfying its internal drives - to respond to the human and to extend its knowledge. We describe different mechanisms that implement different behaviours leading to a coherent compound behaviour that facilitates different kinds of learning initiated by the human tutor or by the system itself. We demonstrate these principles in the case of learning conceptual models of objects and their visual properties.

Executive Summary

In this report we present the final CogX reincarnation of George, the curious robot. It is based on the systems we presented in the previous years; it is, however, more robust and is able of a wider range of behaviours. These behaviours are also better integrated and enable more coherent operation of the system. We increased the robustness of the system by improving the two-layered attention-driven visual subsystem, which is based on more robust RGBD sensor. We fully integrated object learning and recognition into the system, so the robot can now also recognise and talk about object types. In addition, we also reformulated the belief system and introduced merged beliefs as the final result of the information fusion and abstraction; they contain as reliable information about the perceived objects as possible and as much information about them as available. We also better structured the goals the robot is aiming at and introduced three priority levels for drives that generate goals. We assigned the highest priority to the interaction drive as the robot should always try to respond to the human as promptly as possible. On the second level we placed the extrospection drive that generates goals to understand and explore the scene and to learn as much as possible from this information. We assigned the lowest priority to the introspection drive, which tries to improve its models by introspection. This prioritisation together with the planning mechanism is supposed to lead to smooth transitions and appropriate switching between different behaviours leading to an efficient and natural mixed-initiative learning dialogue.

Role of a curiosity driven self-extending robot system in CogX

George tries to understand what it is certain about and what it is not, what it knows and what it does not know. Based on this, it tries to get the missing information (also by interacting with the human tutor) to fill the detected knowledge gaps. Therefore, through curiosity-driven extrospection and introspection the robot tries to extend its current knowledge, which is the major research topic in CogX.

Contribution to the CogX scenarios and prototypes

George is one of three scenarios that we have been addressing in CogX. We have designed this scenario to monitor and show progress on the development and integration of various competencies needed for interactive continuous learning. This scenario has been designed as a use case for guiding and testing system-wide research and for demonstrating methods developed in WP 5, WP 2, WP 1, WP 4, and WP 6 in a working system. Moreover,

George also shares a great part of the code with Dora; the main functionalities in both scenarios are based on the same principles and implementation.

1 Tasks, objectives, results

1.1 Planned work

This deliverable mainly tackles the problems addressed in Task 7.7 of Workpackage 7:

Task 7.7: Integration for full curiosity driven extension system.

As such, it is addressing the following objectives as specified in the Technical annex:

11. *A robotic implementation of our theory able to complete a task involving mobility, interaction and manipulation, in the face of novelty, uncertainty, partial task specification, and incomplete knowledge. [WPs 2,3,6,7]*

12. *Within the same implementation the demonstration of the ability to plan and carry out both task driven and curiosity driven learning goals. [WP 1,7]*

The main goal for the final year of the project was to increase the robustness of the system, as well as to wider the range of different behaviours and to better integrate these behaviours into a coherent compound behaviour. Our objective was to demonstrate that a cognitive system 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.

1.2 Actual work performed

In the last year of the project we substantially extended the George system that was developed in the previous years [6]. We reformulated and re-implemented some of the functionalities (such as the belief system and the prioritisation of the main motivation drives), we added several new functionalities (such as object learning and recognition), and we robustified some of the functionalities (such as attention-driven visual processing), as well as the operation of the system as a whole.

In Annex 2.1 we attach the technical report describing the George robot from the component and from the system point of view; all the individual competencies are briefly described, and also the entire system is shown, focusing on mechanisms that implement different behaviours. It presents how George learns and refines conceptual models of visual objects and their properties, either by attending to information deliberately provided by a human tutor (*Tutor-initiated interaction*: e.g., H: 'This is a Coke can.') or by taking initiative itself. In the latter case, the robot can learn by *extrospection*, i.e., by analysing the objects in the scene and using the acquired

information for updating the knowledge, either automatically, or after asking the tutor for additional information about the objects when necessary, e.g., R: 'Is the elongated object yellow?'. George can also initiate learning by *introspection*, i.e., by analysing its internal models of visual concepts and asking questions that are not related to the current scene, e.g., R: 'Can you show me something red?'. Our approach unifies these cases into an integrated approach including attention-driven visual processing, incremental visual learning, selection of learning goals, continual planning, and a dialogue subsystem. By processing visual information and communicating with the human, the system forms beliefs about the world, which are exploited by the behaviour generation mechanism that selects the actions for optimal learning behaviour. George is therefore a curiosity-driven system that aims at understanding where its own knowledge is incomplete and that takes actions to extend its knowledge subsequently.

The attached technical report will be, when completed, submitted for a journal publication. The journal submission will also include a thorough evaluation of the robot system that is currently being performed and will be completed by the end of the project.

1.3 Relation to the state-of-the-art

In this section we discuss how our work is related to, and goes beyond the current state-of-the-art.

Interactive continuous learning using information obtained from vision and language is a desirable property of any cognitive system, therefore several systems have been developed that address this issue (e.g., [5, 7, 1, 2, 8, 4, 3]). Different systems focus on different aspects of this problem, such as the system architecture and integration [1, 2, 4], learning [5, 7, 4, 3], or social interaction [8]. Our work focuses on the integration of visual perception and processing of linguistic information by forming beliefs about the state of the world; these beliefs are then used in the learning process for updating the current representations. The system behaviour is driven by a motivation framework which facilitates different kinds of learning in a dialogue with a human teacher, including self-motivated learning, triggered by autonomous knowledge gap detection. Also, George is based on a distributed asynchronous architecture, which facilitates inclusion of other components that could bring additional functionalities into the system in a coherent and systematic way (such as navigation and manipulation).

2 Annexes

2.1 Skočaj et al. "An integrated system for interactive learning in dialogue with a tutor"

Bibliography D. Skočaj, M. Kristan, A. Vrečko, M. Mahnič, M. Janíček, GJ M. Kruijff, M. Hanheide, N. Hawes, T. Keller, M. Zillich and K. Zhou: "An integrated system for interactive learning in dialogue with a tutor". To be submitted for journal publication, 2012.

Abstract In this paper we present representations and mechanisms that facilitate continuous learning of visual concepts in dialogue with a tutor and show the implemented robot system. We present how beliefs about the world are created by processing visual and linguistic information and show how they are used for planning the system behaviour with the aim at satisfying its internal drives - to respond to the human and to extend its knowledge. We describe different mechanisms that implement different behaviours leading to a coherent compound behaviour that facilitates different kinds of learning initiated by the human tutor or by the system itself. We demonstrate these principles in the case of learning conceptual models of objects and their visual properties.

Relation to WP The paper describes the final version of the George system, so it is directly related to WP 7. It also briefly describes the individual functionalities of the system that have been developed in other workpackages, namely in WP 5, WP 2, WP 1, WP 4, and WP 6.

References

[1] C. Bauckhage, G.A. Fink, J. Fritsch, F. Kummmert, F. Lomker, G. Sagerer, and S. Wachsmuth. An integrated system for cooperative man-machine interaction. In *In: IEEE International Symposium on Computational Intelligence in Robotics and Automation*, pages 320–325, 2001.

[2] B. Bolder, H. Brandl, M. Heracles, H. Janssen, I. Mikhailova, J. Schmudderich, and C. Goerick. Expectation-driven autonomous learning and interaction system. In *Humanoids 2008. 8th IEEE-RAS International Conference on*, pages 553–560, Daejeon, South Korea, Dec. 2008.

[3] J. de Greeff, F. Delaunay, and T. Belpaeme. Human-robot interaction in concept acquisition: a computational model. In *Development and Learning, 2009. ICDL 2009. IEEE 8th International Conference on*, pages 1–6, June 2009.

[4] Stephan Kirstein, Alexander Denecke, Stephan Hasler, Heiko Wersing, Horst-Michael Gross, and Edgar Körner. A vision architecture for unconstrained and incremental learning of multiple categories. *Memetic Computing*, 1:291–304, 2009.

[5] D. K. Roy and A. P. Pentland. Learning words from sights and sounds: a computational model. *Cognitive Science*, 26(1):113–146, 2002.

[6] Danijel Skočaj, Matej Kristan, Alen Vrečko, Marko Mahnič, Miroslav Janíček, Geert-Jan M. Kruijff, Marc Hanheide, Nick Hawes, Thomas Keller, Michael Zillich, and Kai Zhou. A system for interactive learning in dialogue with a tutor. In *IEEE/RSJ International Conference on Intelligent Robots and Systems IROS 2011*, San Francisco, CA, USA, 25-30 September 2011.

[7] L. Steels and F. Kaplan. AIBO's first words, the social learning of language and meaning. *Evolution of Communication*, 4(1):3–32, 2000.

[8] Andrea L. Thomaz and Cynthia Breazeal. Experiments in socially guided exploration: lessons learned in building robots that learn with and without human teachers. *Connection Science*, 20(2 3):91–110, June 2008.

An integrated system for interactive learning in dialogue with a tutor

Danijel Skočaj, Matej Kristan, Alen Vrečko, Marko Mahnič, Miroslav Janíček, Geert-Jan M. Kruijff, Marc Hanheide, Nick Hawes, Thomas Keller, Michael Zillich and Kai Zhou

Abstract

In this paper we present representations and mechanisms that facilitate continuous learning of visual concepts in dialogue with a tutor and show the implemented robot system. We present how beliefs about the world are created by processing visual and linguistic information and show how they are used for planning system behaviour with the aim at satisfying its internal drives - to respond to the human and to extend its knowledge. We describe different mechanisms that implement different behaviours leading to a coherent compound behaviour that facilitates different kinds of learning initiated by the human tutor or by the system itself. We demonstrate these principles in the case of learning conceptual models of objects and their visual properties.

1. Introduction

Cognitive systems are often characterised by their ability to learn, communicate and act autonomously. By combining these competencies, the system can incrementally learn by engaging in mixed initiative dialogues with a human tutor. In this paper we focus on representations and mechanisms that enable such interactive learning and present a system designed to acquire visual concepts through interaction with a human.

Such continuous and interactive learning is important from several perspectives. A system operating in a real life environment is continuously exposed to new observations (scenes, objects, actions etc.) that cannot be envisioned in advance. Therefore, it has to be able to update its knowledge continuously based on the newly obtained visual information and information provided by a human teacher. Assuming that the information provided

by the human is correct, such interactive learning can significantly facilitate, and increase the robustness of, the learning process, which is prone to errors due to unreliable robot perception capabilities. By assessing the system's knowledge, the human can adapt their way of teaching and drive the learning process more efficiently. Similarly, the robot can take the initiative, and ask the human for the information that would increase its knowledge most, which should in turn lead to more efficient learning.

In this paper we describe how our robot *George*, depicted in Fig. 1, learns and refines conceptual models of visual objects and their properties, either by attending to information deliberately provided by a human tutor (*Tutor-initiated interaction*: e.g., H: 'This is a Coke can.') or by taking initiative itself. In the latter case, the robot can learn by *extrospection*, i.e., by analysing the objects in the scene and using the acquired information for updating the knowledge, either automatically, or after asking the tutor for additional information about the objects when necessary, e.g., R: 'Is the elongated object yellow?'. George can also initiate learning by *introspection*, i.e., by analysing its internal models of visual concepts and asking questions that are not related to the current scene, e.g., R: 'Can you show me something red?'. Our approach unifies these cases into an integrated approach including attention-driven visual processing, incremental visual learning, selection of learning goals, continual planning, and a dialogue subsystem. By processing visual information and communicating with the human, the system forms beliefs about the world, which are exploited by the behaviour generation mechanism that selects the actions for optimal learning behaviour. George is one system in a family of 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 that a cognitive system 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.

Interactive continuous learning using information obtained from vision and language is a desirable property of any cognitive system, therefore several systems have been developed that address this issue (e.g., [1, 2, 3, 4, 5, 6, 7]). Different systems focus on different aspects of this problem, such as the system architecture and integration [3, 4, 6], learning [1, 2, 6, 7], or social interaction [5]. Our work focuses on the integration of visual perception and processing of linguistic information by forming beliefs about the state of the world; these beliefs are then used in the learning process for updating

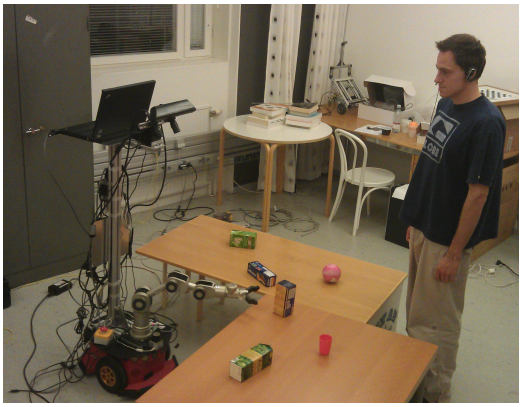

Figure 1: Interactive learning scenario.

the current representations. The system behaviour is driven by a motivation framework which facilitates different kinds of learning in a dialogue with a human teacher, including self-motivated learning, triggered by autonomous knowledge gap detection. Also, George is based on a distributed asynchronous architecture, which facilitates inclusion of other components that could bring additional functionalities into the system in a coherent and systematic way (such as navigation and manipulation).

The paper is organised as follows. In §2 we present the competencies and representations that allow integrated continuous learning. In §3 we describe the system that we have developed and focus on mechanisms that implement different behaviours leading to a coherent compound learning behaviour. An example dialogue is then presented in §4. We conclude the paper with a discussion and some concluding remarks in §5.

2. System competencies and representations

A robotic system capable of interactive learning in dialogue with a human needs to have several competencies (the ones that enable it to demonstrate such behaviour) and has to be able to process the different types of representations stemming from different modalities. Fig. 2 concisely depicts the main competencies of our system and the relationships between them. By processing visual information and communicating with the human, the system forms beliefs about the world. They are exploited by the behaviour generation mechanism that selects the actions to be performed in order to

extend the system's knowledge about visual concepts. In the following we first describe the individual competencies and representations, then show how they are integrated into a unified robot system.

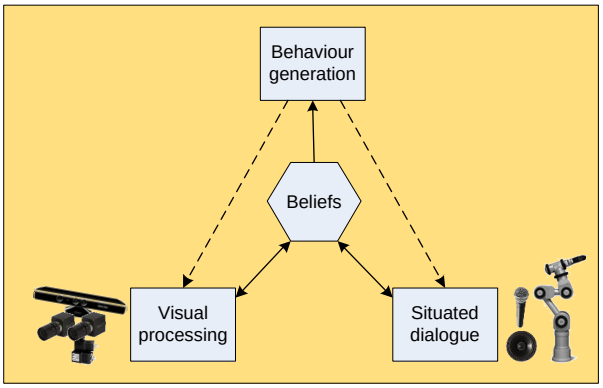

Figure 2: System competencies and relationships between them.

2.1. Attention driven visual processing

To autonomously learn visual object concepts the system needs to identify the moment when new objects are presented as a learning opportunity. Since initially there are no models for these yet, it cannot rely on model-based recognition, but requires a more general mechanism. To this end the system uses a generic bottom-up 3D attention mechanism suited for indoor environments that are typical for many robotic tasks.

To make the problem of generic segmentation of unknown objects tractable we introduce the assumption that objects are presented on a table, or any other supporting surface (which is always the case in the scenario we are addressing). Based on 3D point clouds obtained with an RGBD sensor, the system detects (possibly multiple) supporting planes using a variant of particle swarm optimization [8, 9]. Any parts sticking out from the supporting plane form spaces of interest (SOIs), i.e. anything that is potentially interesting, without regard to its properties. These SOIs are subsequently validated by tracking them over time, based on colour histogram, size and position.

As segmentation based on the RGBD data alone can be imperfect and due to shadowing effects at object boundaries can include points with erroneously assigned background colour, stable SOIs are augmented with a

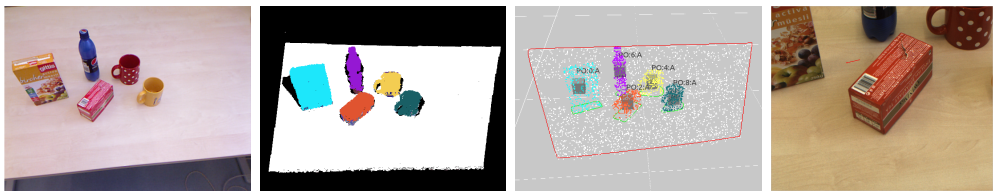

Figure 3: A typical scene after placing 5 objects, with segmented point cloud and resulting proto-objects in working memory, and finally the close-up view of a foveated object.

precise segmentation mask using graph cut [10]. This segmentation happens in a close-up view of the potential object, using a higher resolution RGB image from a camera with a longer focal length than the RGBD sensor. Object properties to be learned, such as colour and shape, are then extracted based on the segmentation mask. Figure 3 shows an example scene with wide angle image, segmented point cloud, extracted proto-objects in working memory and the close-up view after foveating on an object.

2.2. Learning and recognition of object properties

To efficiently store and generalise the extracted visual information, the visual concepts of objet properties, such as colour and basic shapes, are represented as generative models. These generative models take the form of probability density functions (pdf) over the feature space, and are constructed in an online fashion from new observations. The continuous learning proceeds by extracting the visual data in the form of multidimensional features (e.g., multiple 1D features relating to shape, texture, colour and intensity of the observed object) and the *online discriminative Kernel Density Estimator* (odKDE) [11] is used to estimate the pdf in this multi-dimensional feature space. The odKDE estimates the probability density functions by a mixture of Gaussians, is able to adapt using only a single data-point at a time, does not assume specific requirements on the target distribution, and automatically adjusts its complexity by compressing the models. The odKDE penalizes discrimination loss during compression of the generative models that it builds from the data stream, thus introducing a discriminative criterion function in the construction of generative models. A particularly important feature of the odKDE is that it allows adaptation from the positive examples (learning) as well as negative examples (unlearning) [12].

Therefore, during online operation, a multivariate generative model is continually maintained for each of the visual concepts and for mutually ex-

clusive sets of concepts (e.g., all colours) the optimal feature subspace is continually being determined by feature selection. This feature subspace is then used to construct a Bayesian classifier, which can be used for recognition of individual object properties. However, since the system is operating in an online manner, the system could at any moment encounter a concept that has not been observed before. We model the probability of this occurring with an "unknown model", which should account for poor classification when none of the learnt models supports the current observation strongly enough. Having built such a knowledge model and Bayesian classifier, recognition is done by inspecting a posteriori probability (AP) of individual concepts and the unknown model.

Such a knowledge model is also appropriate for detecting gaps and uncertainty in knowledge. By analysing the AP for an object, the system determines the *information gain* for every concept. The information gain estimates how much the system would increase its knowledge, if it were to receive information from the tutor about the particular concept related to a particular object in the scene (e.g., the colour of the object). This serves as a basis for triggering situated extrospective learning mechanisms. Furthermore, the system can also inspect its models and determine which model is the weakest or the most ambiguous. Based on this estimate, the information gain for every concept is again calculated; this time, it does not relate to a particular object and serves as a basis for initiating introspective learning.

2.3. Learning and recognition of object models

Besides generic object properties George also learns individual objects [13]. We use view based 3D object models, learned incrementally and consisting of a series of registered object views, each containing a set of SIFT features together with their 3D position on the object surface. These views are acquired incrementally from RGBD images, and are aligned using sparse bundle adjustment. Recognition then uses RANSAC to find a matching view given SIFT features extracted from a given RGB image. Relying only on RGB images for recognition allows to use a camera with a small field of view for detecting objects further away, where objects would already be too small in the rather large field of view of the RGBD sensor.

Similar to the above object properties the robot maintains not only the models themselves but also measures of their completeness. To this end we define learned probabilistic measures for *observed detection success*, *predicted detection success* and *model completeness*, allowing George to quantify its

current knowledge about an object and the predicted increase in knowledge for a given action (i.e. add a new view after a change of view point). *Observed detection success* $p(o \mid c)$ is the probability of having successfully detected the object o given the detector's confidence value c. Note that while confidence values are often expressed in the range $0 \ldots 1$ they do not constitute actual probabilities. So we have to learn the meaning of that confidence value in terms of actual probability, where we use a series of virtually rendered views of the model acquired so far to generate training examples (see Fig. **??**). *Predicted detection success* $p(o_j \mid \theta)$ is defined as the probability of successfully detecting object view o_j given an out of plane rotation θ, and is again learned using virtual training examples. To arrive at a measure of *model completeness* we take the expected detection probability over all learned views $\hat{p}(o) = \sum_{\theta,j} p(o_j \mid \theta)p(\theta)$ where prior $p(\theta)$ can take into account that certain views are less likely than others and thus possibly not even learned (such as the underside of an object). We then define the knowledge gain g when learning a new view $n+1$ as the expected increase in $\hat{p}(o)$ after learning the new $(n+1)$-th view: $g = \hat{p}_{n+1}(o) - \hat{p}_{n+1}(o)$. I.e. we tentatively add the (empty) future view to our model together with its *predicted detection success* and calculate the increase in detection probability.

We therefore represent the completeness of an object model as the expected detection probability over all object views learned so far, where the detection probability of individual views are learned from virtual training examples. As the model aligns its learned views it can be directly inferred which parts of the view sphere are currently not covered by views and thus represent knowledge gaps, where the knowledge gain measure introduced above quantifies the gain in closing that gap.

2.4. Situated dialogue

In addition to vision, the other source of information for George is dialogue with a human tutor. In task-oriented dialogues between a human and a robot, there is more to dialogue than just understanding words. The robot needs to understand what is being talked about, but it also needs to understand why it was told something. In other words, what the human *intends* the robot to do with the information in the larger context of their joint activity. To do so, we employ continual abduction [14] to generate and verify hypotheses about the human tutor's behaviour in terms of communicative intentions and explicitly represent it in the system.

Abduction is a method of explanatory logical reasoning introduced into modern logic by Charles Sanders Peirce [15]. Given a theory T, a rule $T \vdash A \rightarrow B$ and a fact B, abduction allows inferring A as an explanation of B. B can be deductively inferred from $A \cup T$. If $T \nvdash A$, then we say that A is an *assumption*. There may be many possible causes of B besides A. Abduction amounts to *guessing*; assuming that the premise is true, the conclusion holds too.

Obviously, as there may be many possible explanations for a fact, in practical applications there needs to be a mechanism for selecting the best one. This may be done by purely syntactic means (e.g. lengths of proofs), or semantically by assigning *weights* to abductive proofs and selecting either the least or most costly proof [16], or by assigning probabilities to proofs [17]. In that case, the most probable proof is also assumed to be the best explanation. Our approach combines both aspects.

Intentions are goal-oriented cognitive states usually modelled as distinct from desires in that there is an explicit commitment to acting towards the goal and refraining from actions that may render it impossible to achieve [18, 19]. The communication system explicitly models *communicative intentions*, i.e. intentions that are related to the communication[1], and use them as a pragmatic representation of the human-robot interaction, abstracting away from the actual surface form.

Abductive reasoning over intentions set within a (situated) context is a bi-directional process [20], and is used in our system in two roles:

- recognition of the *tutor's* communicative intentions (given the context and a surface form of the tutor's input, infer his intention),
- realisation of the *robot's* communicative intentions – (given the context and the robot's intention, infer an appropriate surface form).

We employ abduction in a continual manner, explicitly modelling the knowledge gaps that inevitably arise in such an effort due to uncertainty and partial observability. The approach is based on generating partial hypotheses for the explanation of the observed behaviour of other agents, under the assumption that the observed behaviour is intentional. These partial hypothe-

[1] as opposed to, for instance, the robot's purely *internal* intentions that have nothing to do with communication. See also §2.7.

ses are defeasible and conditioned on the validity (and eventual verification) of their assumptions.

The abductive reasoning system represents knowledge gaps as partial abductive proofs. In order to turn partial proofs into "full" proofs, the knowledge gaps in them need to be verified or falsified.

Our extension of the "classical" logic-programming-based weighted abduction as proposed by Stickel and Hobbs [21, 16] lies in the extension of the proof procedure with the notion of *assertion* based on the work in continual automated planning [22], allowing the system to reason about information not present in the knowledge base, thereby addressing the need for reasoning under the open-world assumption.

In continual automated planning, assertions allow a planner to reason about information that is not known at the time of planning (for instance, planning for information gathering), an assertion is a construct specifying a "promise" that the information in question will be resolved eventually. Such a statement requires planning to be a step in a continual loop of interleaved planning and acting.

By using a logic programming approach, we can use unbound variables in the asserted facts in order to reason not only about the fact that the given assertion will become a proven fact, but also under-specify its eventual arguments.

2.5. Modeling beliefs

By processing visual information and communicating with the human, the system forms *beliefs* about the world. Beliefs are data structures that contain indexical information about the perceived entities in the scene. They form a cognitive layer where multi-modal and multi-agent information is associated and merged to a-modal representations. In general a belief can be regarded a high-level representation of an element of the physical reality, grounded in one or more sensory inputs, attributed to a specific agent or a combination of both. Typically, a single belief contains information about one entity, but there can be many beliefs about a single entity. The information inside beliefs is expressed in multivariate probability distributions over feature-value pairs.

An important aspect of the beliefs is their multi-agent aspect: a belief can be private to the robot, attributed to an external agent (e. g. human), or common ground among the robot and one or more other agents. In this sense we distinguish five distinct belief categories:

- *Private* beliefs reflect the robot perceptions of the environment based on its sensory input. Private beliefs are expressed in modal symbols and can form various associations with private beliefs stemming from other modalities or beliefs with other epistemic statuses (e. g. reference resolution).

- *Assumed* beliefs are used to establish cross-agent or cross-modal common ground. They are created from private beliefs by translating the modal symbols to the a-modal ones. Depending on complexity of the modal learners and their ability for autonomous unsupervised learning, this process can be as simple as one-to-one symbol mapping or much more complex (e. g. translating between two sets of symbols with overlapping meaning that consequently also modifies the original probability distribution). In cross-agent case the robot uses assumed beliefs to establish a common ground with another agent to facilitate communication. Thus the beliefs reflect the robot assumptions about the meaning of its perceived information for a particular agent (e. g. human). In cross-modal case the assumed beliefs establish a common ground between modalities. In both cases this process facilitates cross-belief information fusion in later stages.

- *Attributed* beliefs contain information that robot attributes to another agent (e. g. human). This kind of beliefs are the direct consequence of some kind of communication with another agent. The robot is in principle able to analyze and understand the information in such beliefs, but does not necessarily agree with it (especially, if it doesn't match the robot's own perception of the same reality).

- *Verified* beliefs are created from attributed beliefs. They basically contain the acknowledged information from the attributed beliefs. Acknowledgment (verification) does not necessarily mean that the agent's information in the belief is consistent with the robot's perception; it just means that that information was adequately processed by the robot and is now ready to be used in higher level cognition (e. g. in communication with the agent that issued it). After a successful reference resolution the restrictive information is stored in verified shared beliefs, while the asserted information is in attributed belief.

- *Merged* beliefs combine information from verified and assumed beliefs

and represent the final a-modal situated knowledge, ready to be used by the higher level cognitive processes (e.g. motivation, planning). They contain as reliable information as possible and as much information as available. Information can be merged in different ways. E. g. the system can completely trust a certain agent (typically a tutor) so that the merged belief contains all information from the verified belief and only uses the assumed belief to fill the information gaps left by the verified belief. A more complex solution for information fusion involves merging probability distributions over feature values.

The private beliefs are created using the information from the modal subsystems. The attributed and verified beliefs are created as results of successful resolution of another agent's reference. The changes in perception are propagated in real-time through the belief structure from private beliefs to the merged ones. In similar manner the progress in dialogue and dialogue processing (certain events in other subsystems can be treated as acknowledgments for the attributed information) are reflected in changes in attributed and verified beliefs. This means that the process of belief merging is repeated each time new information is propagated to the assumed belief or new attributed information is verified.

2.6. Binding and reference resolution

In [23] we presented a model of *cross-modal binding and learning* system formulated in *Markov logic networks (MLN)*. We call *cross-modal binding* the process of combining two or more modal representations (grounded in different sensory inputs) of the same physical entity into a single multi-modal representation. MLN [24, 25, 26] combine first-order logic and probabilistic graphical models in a single representation. An MLN knowledge base consists of a set of first-order logic formulae (rules) with a weight attached:

$$weight \quad first\text{-}order\_logic\_formula.$$

The weight is a real number, which determines how strong a constraint each rule is: the higher the weight — the less likely that rule is violated. MLN is used to encode the cross-modal knowledge, which is the base for the binding inference.

In George the MLN binding is applied to the belief cognitive layer, where the various beliefs represent perceived and assumed facts that are used to instantiate the rules from the cross-modal knowledge base to the *Markov*

network graphical model. If MLN knowledge represents the general rules encoding relations between concepts (e.g. object properties as color, shape,...), the graphical model encodes the relations between concrete instances (objects) that are currently perceived by the system. A successful inference results in a shared multi-modal representation of a physical entity, also called *binding union*. Successful binding unions can be used as learning samples to improve cross-modal knowledge, i. e. *cross-modal learning*.

In George scenario the binding principles are used for *reference resolution*. Reference resolution is a process akin to binding that tries to associate multi-agent information. In our case the robot uses reference resolution to relate information attributed to a human tutor to its own perceptions, hence it is critical for its ability to make situated dialogue with the human.

MLN are implemented as a special component that process information primarily stored in beliefs. A *MLN engine component* maintains a Markov network graphical model, which makes continuous online inference (MCMC sampling) and can continuously adapt to the changes in the beliefs. MLN engines can also combine the information encoded in the current graphical model with the external information about the correct inference outcome to perform on-line weight learning.

In reference resolution the MLN engine processes information from two distinct sources. Information about perceived entities, which is stored in beliefs, is continuously filtered and fed to the engine. The other source of information is the dialogue subsystem. When the dialogue subsystem recognises a referring expression in the tutor's utterance, it forwards the referring information to the engine. The inference result, which is a probability distribution over perceived entities, is used by the dialogue subsystem to fill in the local knowledge gap in determining the interpretation of the tutor's utterance.

2.7. Motivation and planning

Our system is designed to perform multiple, possibly interleaved, goal-directed activities. For a system that must fill gaps in its own knowledge, it is important that is able to generate and manage its own goals, as the opportunities available to it at runtime may be unknown or unpredictable at design-time. To address this we build on our previous design for a *motivation framework* [27, 28]. This framework encodes the *drives* of the system (the general types of things it wants to achieve) as a collection of *goal generators*, each of which generates particular types of goals for the system based on the

output of the dialogue system plus the current belief state. Each individual goal is a (partial) description of a desired future state for the robot (e.g. one in which it knows the colour of a newly visible object). Before these goals can be *activated*, i.e. made the target of planning and plan execution, they must pass through a management system that selects which of the many possible goals should be pursued by the system. The management step is necessary to allow the robot to prioritise those goals that are more important to it out of all the goals it could possibly achieve.

The goal generators in George create the goals necessary to engage in situated dialogue with a human tutor and to learn about its surroundings. The intention structures produced by the dialogue system to describe utterances made by the human (see §2.4) are monitored by a goal generator. Depending on intention content, this generator creates goals to answer polar or open questions about objects, or to perform tutor-driven learning. Each of these goals contains the address of the single merged belief for the object referenced by the intention, plus additional intention-specific information. An additional generator handles the situation where a collection of related intentions have been generated in response to an ambiguous reference. In this case the goal not only includes the content describing the future state, but an existentially qualified reference to a belief that represents the possible referents of the intention. Part of the planning task is then to resolve this reference. Further goal generators inspect the beliefs created from entries in the visual subsystem, including proto-objects, visual objects and concept models. These create goals to generate visual objects, learn features and improve the model status respectively.

The activation of goals in our system is based on a priority hierarchy of drives. Each level represents a general type of behaviour we have identified that our system should perform. The highest priority drive is to *respond to the human*. This is followed by the drive to fill gaps in knowledge via *extrospection* (i.e. inspecting the world external to the agent). At the lowest level is the drive to fill knowledge gaps by *introspection*. Goals of a particular priority suppress the activation of all goals with lower priorities and are suppressed by all goals with higher priorities. This is accomplished using a simple *attention filter* in our framework. Goals that pass this through this filter enter into the management system. Here they can be ranked according to heuristic information provided by their goal generators and the top ranked goals are activated.

Planning is performed for activated goals on a problem description gen-

erated from the system's belief state. Plan execution, execution monitoring and replanning is managed via a collection of action interfaces which trigger individual components in the modality-specific subsystems. We use the Fast Downward [29] planner, a state of the art planning system based on heuristic forward search. We extended it by a preprocessing routine which enables the support of object fluents and numerical constants by compiling them away, and deal with the uncertainty of the real-world environment by using a continual planning approach [22].

In all three kinds of tasks (answering questions and filling gaps by extrospection or introspection), dialogue with the tutor plays an important role, either in the form of answering or asking questions. Therefore, the planner must find a way to establish common ground with the tutor about the object they discuss. George has two ways to do so: Describing the object verbally, or pointing to it with its arm. As we regard a verbal description as the cheaper one, George will always try to describe the object in question if it has some property that is unique among all objects and where human and robot have already established common ground, and it will choose to use the arm otherwise. Once George and its tutor know which object their discussion is about, the planner determines the correct answer or question, and triggers some learning component if necessary.

3. Integrated system and behaviour mechanisms

3.1. Integrated system

We integrated the competencies described above in a robotic system. The implementation of the robot is based on CAS, the CoSy Architecture Schema [30]. The schema is essentially a distributed working-memory model composed of several subarchitectures (SAs) implementing different functionalities. George is composed of six such SAs, as depicted in Fig. 4 (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 the Kinect RGBD sensor and narrow field-of view Point Grey Flea 2 cameras and identifies spaces of interest, where the potential objects are detected and subjected to individual processing, as described in §2.1. The attention driven visual processing make use also of the Direct Perception pan/tilt unit from the *Spatial SA* for bringing the object of interest into the center of attention.

The beliefs can also be altered by the *Dialogue SA* through dialogue processing. The system uses off the shelf software for speech recognition and production and the developed techniques presented in §2.4 for recognition of human's intentions and realisation of the robot's intentions in the situated context. The robot also uses the Neuronics Katana 6M 5DOF robot arm from the *Manipulation SA* for pointing at the object in the scene to establish a common ground with the tutor.

All of the beliefs are collected in the *Binder SA*, which represents a central hub for gathering information from different modalities (subarchitectures) about entities currently perceived in the environment. They are monitored by the *Planning SA*, which generates the robot behavior as described in §2.7. The beliefs are first used to trigger the motivation mechanism to produce the learning goals and then for generating the planning state. Finally, during execution action requests are sent to the Visual, Spatial, Manipulation, and Dialogue SAs to perform actions that generate the desired behaviour. The actual mechanisms that drive these behaviours are described in the following subsection.

3.2. Basic behaviours

The system is very complex, very heterogeneous, and very integrated. This means that also the basic behaviours require the functionalities implemented in several subarchitectures. And they also require that different functionalities are executed in parallel, but are still kept synchronised. In the following we will briefly describe the mechanisms that implement these different behaviours. These mechanisms are depicted in Fig. 5. Here, the main processing flows are sketched, and only the major components or data structures are emphasised (encircled). Every behaviour is triggered by a particular event in a particular component or data structure; they are marked with a thicker circle.

3.2.1. Mechanisms for visual perception

We will first describe the mechanisms for visual perception, i.e., how George observes the scene. There are two main behaviours that provide the robot with the visual information. The first one is bottom-up driven and is triggered by changes in the scene, assuring that the objects that are brought in the view of the robot are analysed as well as possible. The second one is top-down driven and is triggered by the motivation subsystem and should assure that the robot looks around and analyses the entire scene.

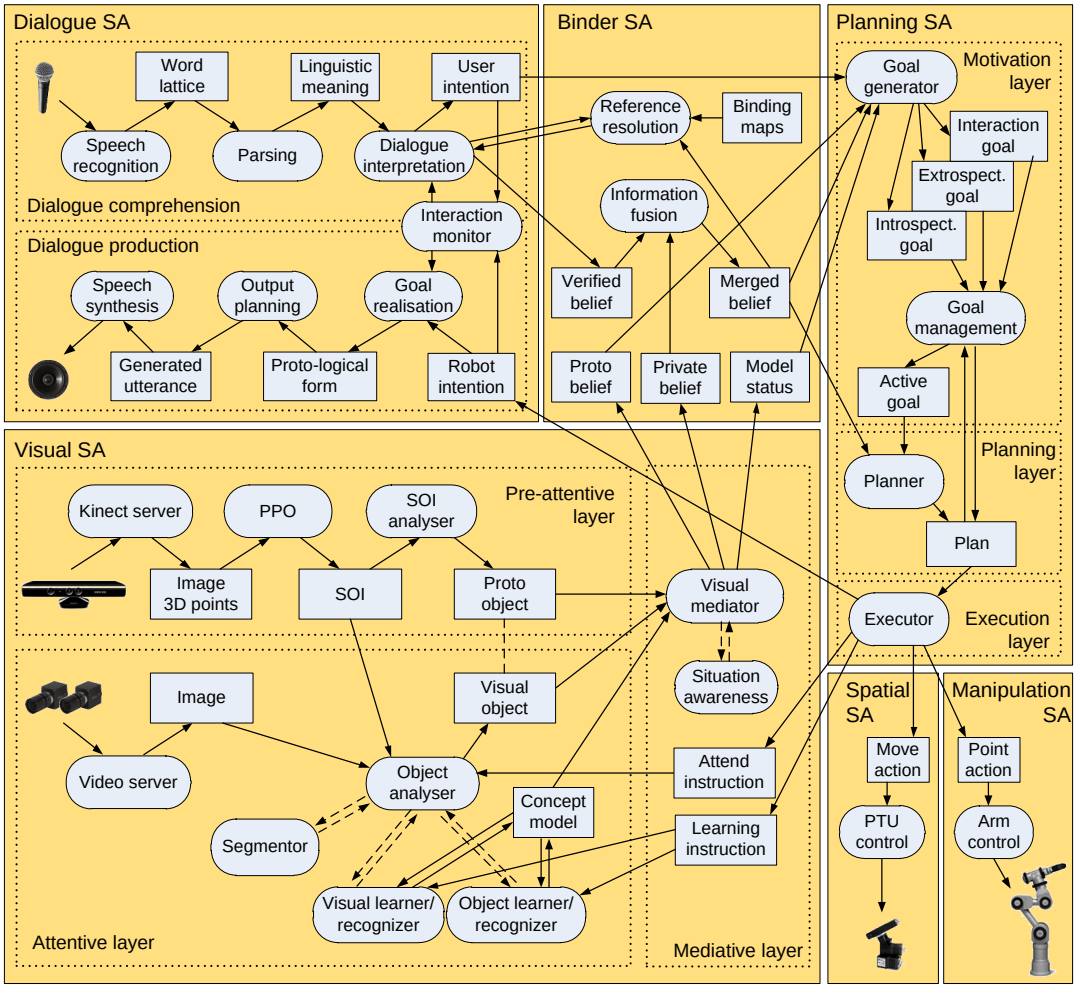

Figure 4: Schematic system architecture.

Both mechanisms are govern by the *extrospection drive* as they relate to understanding the external world.

Attention mechanism. As outlined in §2.1, the most basic behaviour of the system is the bottom-up visual processing based on the plane pop-out attention. Attention leading to generation and tracking of SOIs is always on.

SOIs are transient. So while they are tracked (using colour histogram, size and position) to lead to stable percepts during changes of lighting or small object movements, SOIs are not maintained when George looks away from the scene and back again, and are essentially just the starters for further processing, and not enduring percepts in their own right. Whenever a SOI is found and tracked, the next higher level generates a *proto-object (PO)*, as depicted in Fig. 5(a). A so called view cone is added to the PO, indicating what would be a good close-up look at the PO. The planner detects this goal to get a closer look and prepares a plan for further object analysis. Execution of this plan results in a pan/tilt move to foveate at the new PO, bringing it into the centre of the higher resolution camera, where it will be analysed, i.e. its precise outline will be segmented as an image region of interest (ROI) and object properties (colour, shape) extracted. Furthermore learned object recognisers are run on the ROI in the high-res image, resulting in a label in case the object is already known. All these properties are finally stored in a so called *visual object (VO)*, represented as a *private belief* in the binder. When a newly detected SOI matches an already existing PO (e.g. when the camera moves back to an already analysed scene part) the *visible* attribute of the VO associated with the PO is set from *false* to *true*. Proto-objects thus serve as place-holders for possibly interesting parts of the scene that could become proper visual objects upon further analysis.

Exploring the scene. George has a limited view of the world. There may be objects in front of it but just out of range of its visual system. To make sure George does not miss such objects it has a goal generator which motivates it to move its pan-tilt unit, allowing it to perceive previously unviewed parts of the scene. This generator is triggered after a fixed window of system inactivity, and causes the generation of a small number of view points providing randomly positioned views of the scene. This mechanism is depicted with dashed lines in Fig. 5(a). Clearly, after the robot rotates the cameras to a new orientation, and if in the new view there is a new object, the attention mechanisms described above is triggered and a new object is precessed correspondingly.

3.2.2. Tutor initiated interaction

One of the main capabilities of the system is interaction with a human tutor. The interaction can be triggered by the tutor or by the robot. We first present the mechanisms that govern the interaction initiated by the tutor;

either by asking the robot to execute an instruction or to answer a question, or by giving the robot a useful information that can be used for learning. These mechanisms are triggered by the system's *interaction goals*.

Answering tutor's requests. The communication subsystem is monitoring for the tutor's input at all times. Whenever an input is recognised, its surface form is analysed in terms of the underlying communicative intention. As depicted in Fig. 5(b), after input is detected (either by speech recognition or text input), the word sequence is parsed, assigning a semantic structure to it. This structure is then passed on to the context-sensitive intention recognition module which resolves all references and connects the speech act to the previous discourse. The resulting intention then contains a reference to the object in question, and to previous intentions already present in the working memory. The resulting intention is then written to the working memory, and turned into a motive by the motivation subsystem. The motivation subsystem then decides whether (and when) to make the motive active, in effect reacting to the tutor's input.

From the learning perspective, George can recognise both assertions about the environment (e.g. "The red object is a coke can.") and questions ("What colour is the coke can?", "Is the coke can blue?"). Reacting to the tutor's prompt, on the other hand, is a planned behaviour. This holds both for George's answers to the tutor's questions ("It is blue.") and for George's own initiative ("Could you show me something red?").

The communication subsystem therefore extracts the information about the current scene obtained by the human, and relates this information to the beliefs produced by the robot visual perception. Different beliefs are related and merged in the processes of *reference resolution* (see §2.6) and *information fusion* (see §2.5). In Fig. 5(c) we can see how the reference resolution relates information from the dialogue subsystem (verified beliefs) to perceptual information in private beliefs. Information from both source is then merged by the process of information fusion. The final result of these processes are *merged beliefs* that are used for further higher-level processing.

Situated tutor-driven learning. The tutor can also ask the robot to learn, or provide the information to the robot that can be used for learning. In such situated tutor-driven learning, the learning mechanism is therefore triggered by the tutor. The learning act occurs, when (i) the visual subsystem detects an object and processes its visual features and (ii) the information

provided by the tutor is successfully attributed to the same object. As depicted in Fig. 5(d), this results in the tutor's intention, containing both the reference to the object in question, and encoding the desired effect of the tutor's utterance (i.e., the corresponding change of the private belief about the object). The intention structure is the prerequisite for the motivation subsystem to create a planning goal for visual learning. The goal will be committed to planning and execution only if the expected information gain for the learning action (provided by the visual subsystem) is high enough. Since both prerequisites for the learning are present (visual information from the private belief and a label from the intention), the planner generates a trivial plan – a sequence of learning actions, one for each property provided by the tutor. The execution subsystem delegates the visual learner in the visual subsystem to carry out the actions to update the internal *visual models*, finally resulting in an updated *model status* belief containing key information about the visual models.

3.2.3. Extrospective learning mechanisms

To maximize the learning efficiency a cognitive system has to be able to exploit different kinds of learning opportunities and not only to passively wait for the tutor's learning instructions. It should actively look for, ask for, and use the information that would help to extend its knowledge.

The robot aims at extending its knowledge about visual concepts by minimising the uncertainty about its perceptions of the objects that are currently presented in the scene. In this case, the learning opportunity is represented by a perceived object and by the information available about that object; this information can be obtained by the robot itself or it can be provided by the tutor.

These mechanisms are triggered by the system's *extrospective goals*. The motivation component monitors the *merged beliefs* in the binder; if they contain information that can be exploited for learning, it triggers the learning process.

Situated autonomous learning. If a *merged belief* contains only the information provided by the visual subsystem and this information is reliable enough (therefore the visual concept has been recognised with a high confidence), the motivation triggers the learning cycle. The representations of the corresponding visual concepts are therefore automatically updated, resulting in an updated *model status* belief (as depicted in Fig. 5(e)). In the case of

very confident recognition (with a probability very close to 1), such an update of the knowledge is not necessary, because the current representation can describe the object perfectly well. However, in the case of slightly less reliable recognition, it makes sense to update the knowledge, since it will adapt to the perceived object, and will increase the confidence of the recognition of the same (or similar) objects in the future. However, there is always a danger of incorporating erroneously recognised information into the representations in such an automated way; the system should therefore behave very conservatively and only update the knowledge when the recognition is reliable enough, otherwise it should verify its decision by the tutor.

Situated tutor-assisted learning. The robot can therefore, depending on its current ability to recognise that specific object, ask a question about the object's properties. In this case, the motivation subsystem reacts to the private information in belief only. The robot asks about the object property with the highest *information gain*, since it expects that the model of the corresponding object property will profit most if it gets the information it asks for. In the absence of attributed information the planner generates a more complex plan to ask questions about missing information. The execution subsystem generates a corresponding robot intention, which is further managed by the Dialogue SA, resulting in the synthesis of the corresponding generated utterance (the dashed branch in Fig. 5(e)). Depending on the confidence in the recognition results the planner can select between polar questions (e. g. "Is the color of this object red?") and open questions when the recognition confidence is very low (e. g. "What is the color of this object?"). In the case when the robot can not unambiguously refer to the particular object verbally, it points at it to establish a common ground with the tutor. After the tutor provides the answer, the workflow is similar to the tutor-driven learning.

3.2.4. Introspective learning mechanisms

Even in the absence of situated learning opportunities, the robot can still actively pursue its curiosity motivated goals. E. g. the robot can self-initiatively search for new objects or even ask another agent to show him one (specifying the properties he is currently most interested in). This behavior is based exclusively on the introspection of the existing property models. From a pool of currently maintained property models the robot selects the one that he considers the least adequate (typically inadequately sampled) and based on that initiates an action that tries to obtain new samples to improve it.

Non-situated tutor-assisted learning. The robot tries to obtain new learning samples by making a request to the human tutor (e. g. "Could you show me something red?"). We can see that, as a result of the model instrospection, the robot even tries to influence the quality of the potential new object. The model introspection is performed in the visual subsystem and carried on to the belief layer in the epistemic structure *model status*. The model status has a key role in deciding if and what kind of request to make. It contains key information about the visual models maintained by the visual learner.The most important information is again the *information gain* that in this case estimates the reliability of a model in general, not relating this utility to a particular object in the scene (in contrast, the information gains stored in beliefs denote the utility of new information carried by a particular object).

3.3. Compound behaviour

Very often, different behaviours could be triggered simultaneously, so there is a need for a mechanism that selects among them to assure a coherent compound behaviour. We model this compound behaviour by assigning different priorities to the main drives that raise different goals. The motivation component opts for the goals with higher priorities. Among the goals from the same priority level than the planner selects which one to pursue based on the gains (how much the system is expected to benefit if the goal is fulfilled) and the costs of the actions. The information about the gains is stored in the beliefs and is based on an analysis of the models of the visual concepts and objects that are currently present in the scene.

Table 1 lists three main drives that trigger the behaviours described above. The interaction drive has the highest priority, since we want that the robot reacts to tutor's assertions or requests promptly; this is a basic requirement for a natural robot-tutor dialogue.

Goals to explore the scene and to learn as much as possible about the objects currently presented in the scene are part of the extrospection drive as they relate to understanding the external world. They will be suppressed by interaction goals but at the same time they are prioritised above model introspection. The robot first tries to learn as much as possible about the objects in the current view by attending them and updating the knowledge based on the obtained information. When these goals are not active any more, the exploration behaviour is triggered to explore the wider scene as it can yield new objects which can be learnt about.

Table 1: Priority levels

.

Interaction drive	Answering tutor's requests Situated tutor-driven learning
Extrospection drive	Attention mechanism Situated autonomous learning Situated tutor-assisted learning Exploring the scene
Introspection drive	Non-situated tutor-assisted learning

Non-situated tutor-assisted learning is triggered by goals from the introspection drive at the lowest priority level. As such it is only carried out when no other goals are active (i.e. when all visible objects have had their properties learnt and no scene exploration is necessary). Therefore, when the robot doesn't have anything else to do, it asks the tutor to show it an object with particular visual properties that would potentially increase the robot's models of these properties most.

We chose this particular drive prioritisation to reflect the desired behaviour of the robot: it should always try to respond to the human, then try to understand the scene in front of it (as this will be the subject of future interactions), then try to understand the world in more general terms (e.g. through improving its models).

4. Example dialogue

A good way of describing the behaviour of the developed system is to present a sample dialogue between the robot and the human tutor during learning of visual concepts, such as colour, shape and object models. The robot is asked to recognize and describe the objects in a table top scene, of which there are up to five. The human can move or remove objects from the table during the dialogue, and teach the robot about the objects by describing them. Initially the tutor drives the learning, but after a while, the robot takes the initiative, and is able to learn either without verbal feedback, or by asking the tutor for clarification when necessary. To achieve this the robot must establish a common understanding with the human about what is in the scene, and verbalize both its knowledge and knowledge gaps. In a dialogue

with the tutor, the robot keeps extending and improving the knowledge. To test what the robot has learned the tutor asks questions about the scene. The goal of learning is for the robot's representations to be rich enough to correctly describe the scene.

Consider an empty scene. The tutor puts an object and the robot looks at it by applying the *attention mechanism*.

H: Do you know what this is?

R: No.

At the beginning the robot knows nothing about the objects. *Situated tutor-driven learning* is therefore suitable during these initial stages, since the robot has to be given information to reliably initiate its visual concepts.

H: This is a red object.

R: Let me see. OK.

After George gets this information, it can initiate its visual representation of redness. After several such learning steps, the acquired models become reliable enough that they can be used by George to refer to individual objects, and to understand references by the human. From this point on there can be several objects in the scene at the same time, and by applying the mechanism for *answering tutor's requests* George can understand and answer questions about some of them:

H: What colour is the coke can?

R: It is red.

When enough of the models are reliable, George can take the initiative and drive the learning by asking questions of the tutor. It will typically do this when it is able to detect an object in the scene, but is not certain about some or all of its properties. In such *situated tutor-assisted* learning there are two types of uncertainty and gaps. If the object does not fit any previously learned models, the robot considers there to be a gap in its knowledge and asks the tutor to provide information about its novel property:

R: Which colour is this object?

H: It is yellow.

R. OK.

The robot is now able to initialize the model for yellow and, after the robot observes a few additional yellow objects, which make the model of yellow reliable enough, it will be able to recognize the yellow colour.

In the second case, the robot is able to associate the object with a particular model, however the recognition is not very reliable. Therefore, the robot asks the tutor for clarification:

R: Is this red?

H: No. This is yellow.

R. OK.

After the robot receives the answer from the tutor, it corrects (unlearns) the representation of the concept of red and updates the representation of yellow.

In a similar case as above, but if the recognition of an object is more reliable, George updates the models without asking a question utilising the mechanism for *situated autonomous learning*. Since there is no verification from the tutor, George would be very conservative when using this learning mechanism; it would update the knowledge only when it is well certain about the recognition result.

In this way George attends and processes all the objects it detected in the scene. It therefore knows everything about these objects. It then *explores the scene* by looking around to see if there is any new object. Let us assume that there is no one.

Since there is no new object it could learn about, it tries to detect a gap in its knowledge by introspection. By using the mechanism for *non-situated tutor-assisted learning* it asks the tutor to show it an object, which would help it to improve its knowledge.

R: "Please, show me something blue."

H: "This is a blue object."

R: "Thank you."

The robot can refer to objects verbally or by pointing. Let's say that the tutor empties the table and then puts two objects in the scene. George detects and attends both. If it can recognise one object attribute, which is not equal for both the objects in the scene, it can refer to that object verbally.

R: "Is the Pepsi can blue?"

H: "Yes, it is."

R: "Thank you."

However, if it can not distinguish between the two objects using verbal description, it refers at an object by pointing at it.

R points at one object. R: "What colour is this object?"

H: "It is yellow."

R: "Thank you."

In the same way, George can also choose between using verbal description or pointing for clarification. The tutor removes the objects and puts two other objects in the scene. George recognizes two yellow objects (with a similar probability) and does not reliably recognize the type, therefore it has

to point at an object to clarify the situation.

H: "What shape is the yellow object?"

R points at one object and asks: R: "Do you mean this one?"

H: "Yes."

R: "It is elongated."

Then the tutor removes the objects and puts another two objects in the scene. The robot recognizes two yellow objects and also recognizes the type (a tennis ball and a banana). Now, the robot can clarify the situation verbally by using visual attributes.

H: "What shape is the yellow object?"

R: "Do you mean the tennis ball?"

H: "No."

R: "Banana is elongated."

In such mixed-initiative dialogue George continuously improves its representations of basic visual concepts. After a while George can successfully recognize the acquired concepts to provide reliable answers to a variety of questions.

H: Do you know what this is?

R: It is a blue object.

H: What is the green object?

R: It is an ice tea bottle.

H: What shape it is?

R: It is elongated.

5. Conclusion

In this paper we presented representations and mechanisms that facilitate continuous learning of visual concepts in dialogue with a tutor and showed the implemented robot system. We briefly presented how the beliefs about the world are created by processing visual and linguistic information and how they are used for planning the system behaviour with the aim of satisfying its internal drives – to respond to the human and to extend its knowledge. We focused on different mechanisms that implement different behaviours leading to a coherent compound learning behaviour. We demonstrated these principles in the case of learning conceptual models of objects and their visual properties.

During our research, we have made several contributions at the level of individual components, as well as at the system level. In this paper we wanted to show how an integrated approach comprising attention-driven visual processing, incremental visual learning, selection of learning goals, continual planning to select actions for learning behaviour, and a dialogue subsystem, can lead to a coherent and efficient system capable of mixed-initiative learning. Such an integrated robotic implementation enables system-wide research and development and testing on the system and sub-system level.

The robotic implementation is based on a distributed asynchronous architecture, which facilitates inclusion of other components that will bring additional functionalities into the system in a coherent and systematic way, such as navigation and manipulation. This will increase the possibilities of interaction with the environment and enable the robot to acquire novel information in an even more active and autonomous way. Here, the detection of knowledge gaps and planning for actions that would help to fill these gaps will play an even more important role and will enable more autonomous and efficient robot behaviour. The presented behaviour generation mechanism is general enough to accommodate also such new types of behaviours.

Building on this system, our final goal is to produce an autonomous robot that will be able to efficiently learn and adapt to an ever-changing world by capturing and processing cross-modal information in an interaction with the environment and other cognitive agents.

References

[1] D. K. Roy, A. P. Pentland, Cognitive Science 26 (2002) 113–146.

[2] L. Steels, F. Kaplan, Evolution of Communication 4 (2000) 3–32.

[3] C. Bauckhage, G. Fink, J. Fritsch, F. Kummmert, F. Lomker, G. Sagerer, S. Wachsmuth, in: In: IEEE International Symposium on Computational Intelligence in Robotics and Automation, pp. 320–325.

[4] B. Bolder, H. Brandl, M. Heracles, H. Janssen, I. Mikhailova, J. Schmudderich, C. Goerick, in: Humanoids 2008. 8th IEEE-RAS International Conference on, Daejeon, South Korea, pp. 553–560.

[5] A. L. Thomaz, C. Breazeal, Connection Science 20 (2008) 91–110.

[6] S. Kirstein, A. Denecke, S. Hasler, H. Wersing, H.-M. Gross, E. Körner, Memetic Computing 1 (2009) 291–304.

[7] J. de Greeff, F. Delaunay, T. Belpaeme, in: Development and Learning, 2009. ICDL 2009. IEEE 8th International Conference on, pp. 1–6.

[8] K. Zhou, A. Richtsfeld, M. Zillich, M. Vincze, in: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2011).

[9] K. Zhou, K. M. Varadarajan, M. Zillich, M. Vincze, in: IEEE International Conference on Robotics and Biomimetics (ROBIO), Phuket, Thailand.

[10] Y. Boykov, O. Veksler, R. Zabih, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (2001) 1222–1239.

[11] M. Kristan, A. Leonardis, in: International Conference on Pattern Recognition, Istanbul, Turkey, pp. 581–584.

[12] M. Kristan, D. Skočaj, A. Leonardis, Image and Vision Computing 28 (2010) 1106–1116.

[13] M. Zillich, J. Prankl, T. Mörwald, M. Vincze, in: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).

[14] M. Janíček, in: M. Slavkovik, D. Lassiter (Eds.), New Directions in Logic, Language, and Computation, Springer, 2012.

[15] K. T. Fann, Peirce's Theory of Abduction, Mouton, The Hague, The Netherlands, 1970.

[16] M. E. Stickel, Annals of Mathematics and Artificial Intelligence 4 (1991) 89–105.

[17] D. Poole, Artificial Intelligence 64 (1993) 81–129.

[18] M. Bratman, Intentions, Plans, and Practical Reason, Harvard University Press, Cambridge, MA, USA, 1987.

[19] P. R. Cohen, H. J. Levesque, Artificial Intelligence 42 (1990) 213–261.

[20] M. Stone, R. H. Thomason, in: Proceedings of DIABRUCK 2003: 7th workshop on the semantics and pragmatics of dialogue.

[21] J. R. Hobbs, M. E. Stickel, D. E. Appelt, P. A. Martin, Artificial Intelligence 63 (1993) 69–142.

[22] M. Brenner, B. Nebel, Journal of Autonomous Agents and Multiagent Systems (2009).

[23] A. Vrečko, A. Leonardis, D. Skočaj, Neurocomputing (2012) In press.

[24] M. Richardson, P. Domingos, Mach. Learn. 62 (2006) 107–136.

[25] P. Domingos, Data Min. Knowl. Discov. 15 (2007) 21–28.

[26] P. Domingos, M. Richardson, in: Proc. of the ICML-2004 workshop on statistical relational learning and Iits connections to other fields, pp. 49–54.

[27] N. Hawes, Artificial Intelligence 175 (2011) 1020–1036.

[28] J. L. Wyatt, A. Aydemir, M. Brenner, M. Hanheide, N. Hawes, P. Jensfelt, M. Kristan, G.-J. M. Kruijff, P. Lison, A. Pronobis, K. Sjöö, D. Skočaj, A. Vrečko, H. Zender, M. Zillich, IEEE Transactions on Autonomous Mental Development 2 (2010) 282 – 303.

[29] M. Helmert, Journal of Artificial Intelligence Research 26 (2006) 191–246.

[30] N. Hawes, J. Wyatt, Adv. Eng. Inform. 24 (2010) 27–39.

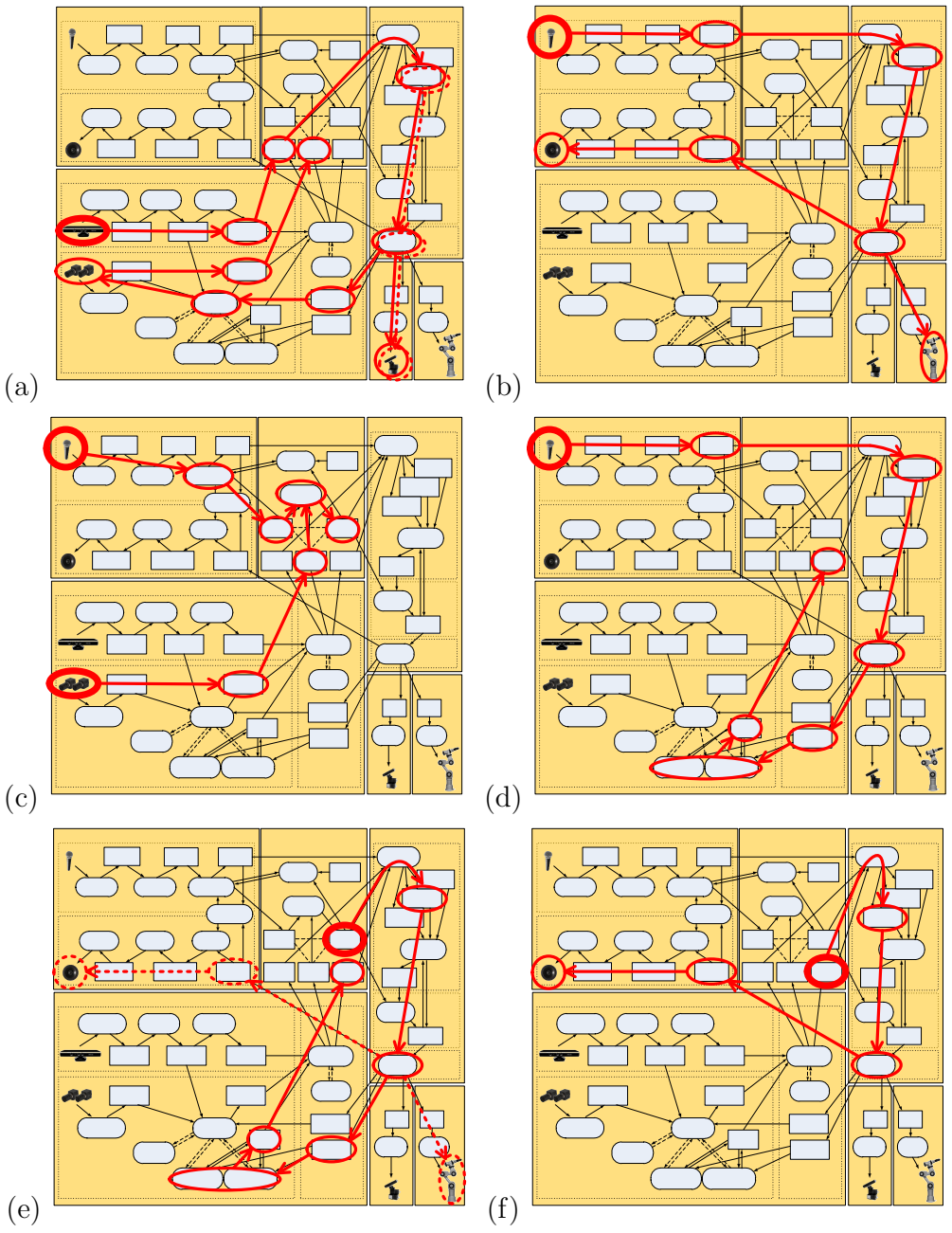

Figure 5: Behaviour mechanisms. (a) Attention and exploration mechanisms. (b) Answering tutor's requests. (c) Merging multi-modal information. (d) Situated tutor-driven learning. (e) Situated autonomous and tutor-assisted learning. (f) Non-situated tutor-assisted learning.