A 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 drive – to extend its knowledge. The system facilitates different kinds of learning initiated by the human tutor or by the system itself. We demonstrate these principles in the case of learning about object colours and basic shapes.**

I. 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 visual conceptual models of colours and two basic shapes, either by attending to information deliberately 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 about an object in the scene (*situated tutor-assisted learning*: e.g., G: 'Is the elongated object yellow?'), or even asking questions that are not related to the current scene (*non-situated tutor-assisted learning*: e.g., G: 'Can you show me something red?')[1]. Our approach unifies these cases into an integrated approach including 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 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.

Fig. 1. Scenario setup.

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 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,

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

D. Skočaj, M. Kristan, A. Vrečko, and M. Mahnič are with University of Ljubljana, Slovenia

M. Janíček and G.J. M. Kruijff are with DFKI, Saarbrücken, Germany

M. Hanheide and N. Hawes are with University of Birmingham, UK

T. Keller is with Albert-Ludwigs-Universität Freiburg, Germany

M. Zillich and K. Zhou are with Vienna University of Technology, Austria

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

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 §II we present the competencies and representations that allow integrated, continuous learning, and describe the system we have developed. In §III we focus on different types of learning mechanisms. The experimental results are then presented in §IV. We conclude the paper with a discussion and some concluding remarks in §V.

II. 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 properties. In the following we first describe the individual competencies and representations, then show how they are integrated into a unified robot system.

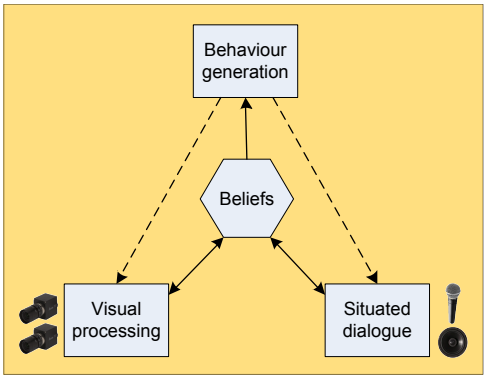

Fig. 2. System competencies and relationships between them.

A. Vision

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. Based on 3D point clouds obtained from a stereo rig, the system detects (possibly multiple) supporting planes using a variant of particle swarm optimization [8]. 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 persistence, stability and size.

As segmentation based on the stereo 3D point cloud alone tends to be imperfect and can include background, especially for weakly textured objects, stable SOIs are augmented with a precise segmentation mask using the graph cut algorithm [9] based on combined colour and 3D information. Object properties to be learned, such as colour and shape, are then extracted based on the segmentation mask.

B. Visual learning and recognition

To efficiently store and generalise the extracted visual information, the visual concepts 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) [10] 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) [11].

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) 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 tutor-assisted learning. 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 non-situated tutor-assisted learning.

C. Beliefs

Each unit of information describing an entity (e.g., an object) is expressed as a probability distribution over a space of alternative values (e.g., different colours, or different shapes). These values are formally expressed as propositional logical formulae. The resulting system is given formal semantics by translating the units of information into formulae in Markov Logic [12]. We call these units of information *beliefs* [13].

Beliefs are constrained both *spatio-temporally* and *epistemically*. They include a frame stating where and when the described entity is assumed to exist, and an epistemic status stating for which agent(s) (the robot, the human tutor) the information contained in the belief holds. Finally, beliefs are also given an *ontological category* used to sort the various belief types.

The *epistemic status* of an epistemic object indicates for which agent(s) the information in the object holds. We define three epistemic statuses of beliefs:

- *Private beliefs*, coming from within the robot as direct or indirect results of its experience of the environment.
- *Attributed beliefs*, i.e. beliefs about the human's beliefs, are the robot's conjecture about the cognitive state of the human tutor. These are typically an indirect result of intention recognition (language understanding).
- *Shared beliefs*, denoting the robot's view of the common ground between the robot and the human.

Besides beliefs, which represent situated information, other kinds of epistemic objects are needed for nonsituated information, e.g. information gathered by the system over several entities, but not specifically tied to any of them. One such type of epistemic object, representing models for modal concepts (e.g. generative models for visual properties, see II-B), is called a *model status*.

Beliefs, being high-level symbolic representations, provide a shared model of the environment which can be therefore altered by dialogue and further exploited by higher-level processes such as motivation and planning.

D. Situated dialogue

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.

Therefore, understanding language can be phrased as an *intention recognition* problem: given an utterance from the human, how do we find the intention behind it? We extend Thomason and Stone's abductive account of language understanding, planning and production [14], in which agents actively monitor and maintain common ground, and to this end they attempt to abductively recognize the others' *intentions* as explanations of their observed (linguistic) behaviour. Our extension of this approach is based on explicit reasoning over the beliefs of agents involved in the interaction [15].

Conceptually, we can distinguish three main components in charge of the robot's language competence:

- *Language understanding*, i.e. the process of recognising the intention behind the human's utterance. This includes relating linguistic expressions such as references to entities in the situated (belief) context.
- *Dialogue management* is a deliberative component in the situated dialogue loop. Given a context update (e.g. a recognised intention), dialogue management selects actions to be performed by the language subsystem. This action is also expressed as an intention, this time the *robot's* intention to act.
- *Language production* is then the process of realising the robot's intention given the situated context.

E. Behaviour generation

In order to create intelligent behaviour an integrated collection of competencies, systems such as George require mechanisms to marshall these competencies, in pursuit of desired future states. For reasons of generality and flexibility we have chosen to use *AI planning* to generate intelligent behaviour in George. There are three elements to planning which must be tightly integrated in an intelligent robot: goal generation and management; planning; and execution. Execution in our system is relatively simple (a set of mediator components that trigger other components when a plan requires it), so here we will focus on the two preceding steps in the process.

For an intelligent robot to be truly autonomous it must be capable of generating its own goals and selecting which ones to pursue when [16]. George features a *motivation framework* which is capable of generating goals from the results of sensing and internal processing, and selecting which of these to pass on to planning. Goals are generated to satisfy *drives*, general dispositions to attain particular future states. George has one primary drive: to extend its knowledge. This drive leads George to be *curious* about its world. We have previously shown the benefits of a motivation framework featuring such a drive in a mobile robot [17] and are now exploring its use in learning and dialogue. The knowledge extension drive has three associated goal generators. The first generates goals for learning when the human provides tutoring information about an object, and a corresponding attributed belief is created (tutor-driven learning). The second goal generator monitors the private beliefs of the robot for perceived objects. If any object belief indicates that there is uncertainty in the underlying representation of the corresponding concept (colour or shape), a goal is generated to ask the human for information about this property (situated

tutor-assisted learning). The final goal generator inspects the model status, an epistemic object carrying the information about the learnt models of visual concepts. It generates goals to ask to see new objects with particular properties if the models for these properties are not particularly discriminative (non-situated tutor-assisted learning).

The motivation framework selects which goal to achieve based on their potential information gain and associated cost. These values are derived from the system's models and the reliability of recognition of the currently observed objects, and are stored in the beliefs. The selected goals are forwarded to the planning subsystem.

At the heart of the planning subsystem is Fast Downward [18], a *classical* planner. Given an initial state, a set of actions, and a goal formula, classical planning is about finding sequences of actions turning the initial state into a state satisfying the goal formula. As the classical planning approach relies on having a complete and certain description of the situation the agent is faced with, a condition that is not met in the George scenario, we extend the planner to handle uncertainty using *continual planning* [19]. In this *optimistic* approach, the planner assigns desired effects to actions with uncertain outcomes, and *monitors* their execution in order to *replan* whenever the optimistic assumption was violated. Combined with the described goal generation and selection processes this results in a robust, domain-independent and easily expandable system that controls the robot's behaviour.

F. The 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 [20]. 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. 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 stereo pairs of images and identifies spaces of interest, where the potential objects are segmented and subjected to individual processing, as described in §II-A. Fig. 3 depicts a sample observed scene and segmented 3D points as well as detected objects. The visual features are then extracted, and used for recognition (and learning) of objects and qualitative visual attributes using the methodology outlined in §II-B. Based on the recognition results, a private belief about every object is generated.

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 §II-D for recognition of human's intentions, reference resolution, and realisation of the robot's intentions in the situated context.

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 subsection II-E. 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 and the Dialogue SAs to perform actions that generate the desired behaviour. The actual mechanisms that drive these behaviours are described in the following section.

Fig. 3. Observed scene and detected objects.

III. Learning mechanisms

To maximize learning efficiency a cognitive system has to be able to exploit different kinds of learning opportunities that require different kinds and levels of learning initiative. In our case a learning opportunity is represented by a perceived object and by the information available about that object, while the learning initiative (besides the learning act itself) involves acquiring new information about the perceived object from the tutor. In this sense we designed three approaches for obtaining required information from the tutor. All three approaches can be used in combination, i.e. in mixed-initiative learning (dialogue). These learning mechanisms are described in the following subsections; the most important part of the process-flow for each of them is also depicted in Fig. 4.

A. Tutor-driven learning

In the tutor-driven learning mechanism, the robot relies on the tutor's initiative to provide information about the visible objects. 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. This results in two beliefs in the binder subsystem: a private belief about the object and the recognized object properties for the visual information; and an attributed belief about the same object for the information provided by the tutor. These two beliefs are the prerequisites 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

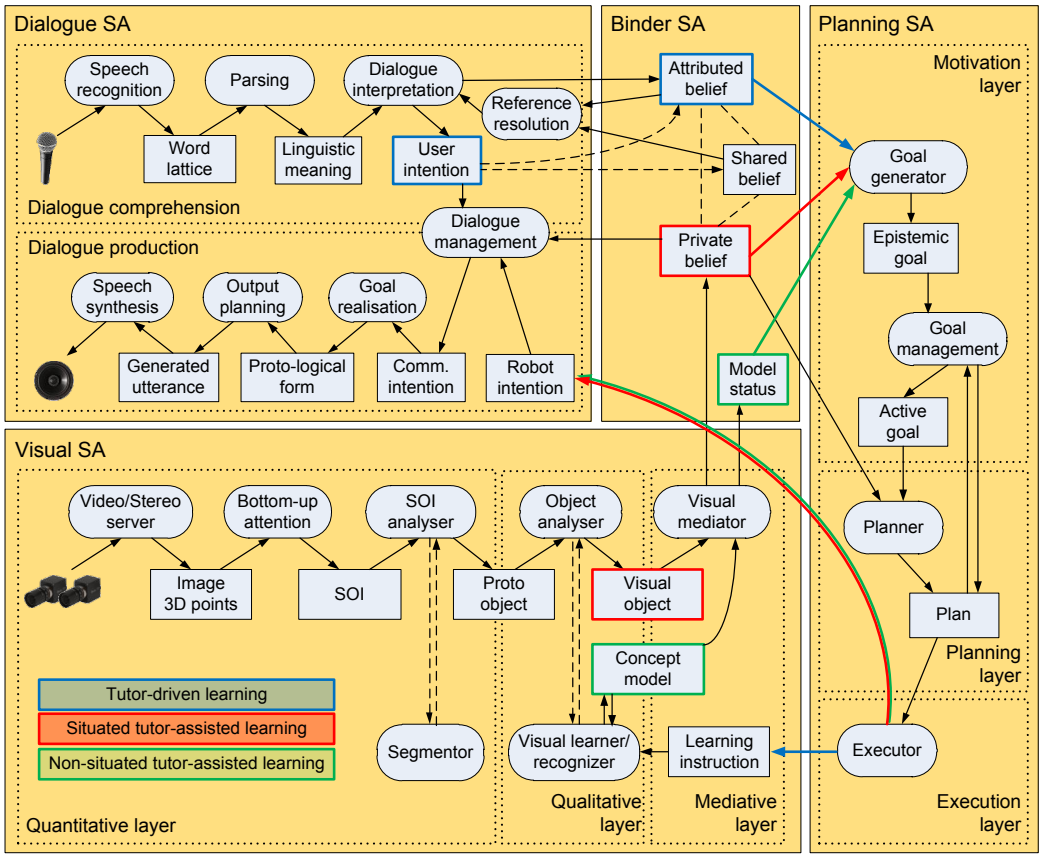

Fig. 4. Schematic system architecture with indicated process flow for three learning mechanisms.

the attributed belief), 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.

B. Situated tutor-assisted learning

In situated tutor-assisted learning the robot shows a greater degree of initiative. In fact, if the tutor does not provide information about a visible object, the robot can, 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 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 an attributed belief 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. 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?"). After the tutor provides the answer, the workflow is similar to the tutor-driven learning.

C. Non-situated tutor-assisted learning

The robot's initiative goes even a step further in non-situated tutor-assisted learning. Here the robot also tries to influence the visual information it is getting by making requests to the tutor (e. g. "Could you show me something red?"). The *model status* (an epistemic structure similar to a belief) has a key role in deciding if and what kind of request to make. The model status contains key information about the visual models (knowledge) maintained by the visual learner. The most important information is again the *information gain*, which 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 gain stored in the private belief denotes the utility of new information carried by a particular object). The goal generator that triggers this learning mechanism has the lowest priority and is usually triggered when no objects are present in the scene. 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.

D. Sample dialogue

A sample, mixed-initiative dialogue is shown in Fig.5. The first fragment shows the dialogue at the beginning, when

the system has not formed any model of any concept yet, so it cannot recognize the object properties. The tutor then presents individual objects to the robot and describes them (tutor-driven learning, *TD*, in Fig.5). After the system gets the information, it can first initiate and later sequentially update the corresponding information. After a number of such learning steps, the acquired models become more reliable and George can take the initiative and try to learn without being told to. In this curiosity-driven learning George can pose an open or a polar question to the tutor, when it is able to detect the object in the scene, but is not certain about their recognition (tutor-assisted learning, *TA*). After the tutor provides an answer, the robot updates or unlearns the corresponding representations. Furthermore, the robot can also analyse its internal models and ask the tutor to show an object of a particular object property that would potentially improve its representations most, not relating its questions to the particular scene (non-situated tutor-assisted learning, *NS*). In such mixed initiative dialogue, George continuously improves the representations and learns reliable models of basic visual concepts. After a while George can successfully recognise the acquired concepts and provide reliable answers.

H: Hi, Robot.
R: Hi, there.
H places an unknown object in front of the robot:
Do you know what this is?
R: I don't know.

H: This is an elongated object.
R: Let me see.
R verifies the assertion and modifies its model: OK.
TD

R: What colour is this object?
H: It is blue.
R: OK.
TA

H: Is this object red?
R: No, it is blue.
H places a green ball on the table:
Could you tell me what shape it is?
R: It is compact.
H: Is the compact object green?
R: Probably yes.

H places another object on the table.
R: Is this object red?
H: No. It is blue.
R: OK.
TA

H removes the objects from the table.
R: Would you show me something yellow?
H places a banana in front of the robot:
This is yellow.
R: I see. Is it also elongated?
H: Yes, it is.
OK.
NS

H places several objects on the table: What colour is the compact object?
R: It is orange.
H: Thanks.
R: Good bye.

Fig. 5. Fragments from a mixed-initiative dialogue.

IV. Experimental results

A. System evaluation

Learning is a long-term phenomena; the system needs to observe a number of training samples to build a reliable and robust representation of the concept that is being modeled. And in order to obtain reliable results, the experiment has to be repeated several times. To test learning performance it is therefore very impractical to conduct an experiment with a real human tutor showing and describing the objects to the system live. Instead, we captured the pairs of images and manually labelled them. Then, we replaced the image stream coming from the live cameras by reading these images from file. We also implemented a simple finite automata that emulated the tutor behaviour in the case of the tutor-driven learning; since the ground truth information about the visual properties of the objects was known, the emulated tutor could describe every image that was shown to the system. Apart from the camera input and speech recognition, the entire system worked in the same way as in the case of live operation, therefore we were able to evaluate the performance of the whole system.

We collected a database of 1120 images of 129 objects; some of them are shown in Fig. 6. We used 500 pairs of images as training samples and the rest of them for testing the recognition performance. The training images were shown to the system one by one and the emulated tutor provided the corresponding description of the objects' properties, which triggered the learning action, as in the case of tutor-driven learning described in §III-A. Eight colours (red, green, blue, yellow, black, white, orange, pink) and two basic shapes (compact, elongated) were being taught. After each update we evaluated the models by trying to recognize the colours of the objects in all test images. The model performance was evaluated in terms of recognition rate. We repeated the experiment three times by randomly splitting the set of images into training and test sets and averaged the results across all runs.

Fig. 6. Sample objects.

The experimental results are shown in Fig. 7. It shows the evolution of the learning performance over time. It is evident that the recognition rate improves with increasing numbers of observed images. The growth of the recognition rate is very rapid at the beginning when new models of newly introduced concepts are being added, and still remains positive even after all models are formed due to refinement of the corresponding representations. The growth is not strictly monotonic; some updates may also cause a drop in the recognition performance due to restructuring of the models, or, more problematically, due to a bad segmentation of the object, which may lead to poor feature extraction. Eventually, by observing additional training samples the models get improved and recognition performance grows again.

At the beginning of the experiment the system knew nothing about the object properties, while at the end it was able to successfully recognize almost all of them. The final recognition rate is 92.40% on average. Most of the misclassifications are due to soft (or even ambiguous) borders between certain colours (such as orange – yellow, pink – red, dark

blue – black). In fact, we asked 10 people to label the same database, and their results differ in a very similar way. Their average recognition rate with respect to the labels that were used to train the robot system is 93.27%. These experimental results show that the entire system performs as expected; it is able to successfully detect the objects, understand the tutor describing these objects and build reliable models of visual properties.

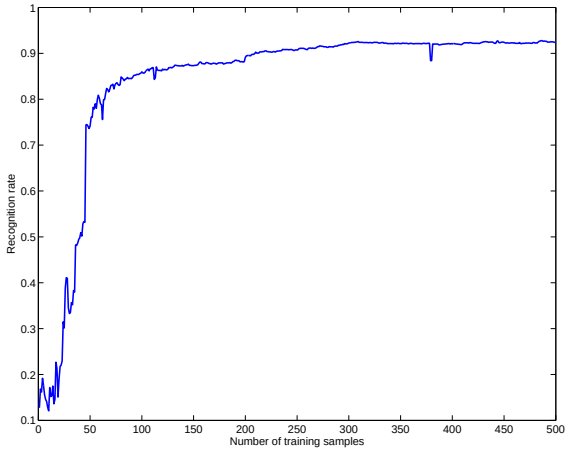

Fig. 7. System evaluation - recognition rate.

B. Evaluation of learning mechanisms

To test the other learning mechanisms we would have to implement significantly more advanced tutor emulation (in fact, we would have to implement another Dialogue SA that would understand the robot's utterances), therefore we performed the evaluation of the proposed learning mechanisms in a simulated environment in Matlab.

We used the same set of 1120 pairs of images as in the previous experiment. We ran the visual subsystem of George, which was used for detecting and segmenting the objects. The extracted features were then used for evaluation of the learning mechanisms.

We tested the performance of all three learning mechanisms presented in §III: tutor-driven (TD), situated tutor-assisted (TA) and non-situated tutor-assisted (NS) learning. In the tutor-driven case we also wanted to test the influence of order of training samples, so we evaluated two variants of the tutor-driven strategy. In the first case (TD_{rnd}), the training images were randomly chosen, while in the second case (TD_{seq}) the models were first initialized with five images from every class and then the objects were presented in a sequence by presenting all objects of the first class, then the second and so on. In both cases the simulated tutor provided the label for every training image. In the case of TA learning, the tutor was randomly presenting the objects, but did not label them. The robot inspected the colour of an object and if it could partially recognize the colour, it would ask the tutor if the recognised colour label was correct. The tutor would answer either "yes" or "no"; in the latter case it would also provide the correct colour label. In the case of NS learning, the robot dictated the sequence in which the training samples were presented by inspecting its internal knowledge and asking the tutor to present an object of a particular colour.

We evaluated the performance of the learned models in terms of the recognition rate obtained on the training set. However, in such interactive learning settings, the success of recognition is not the only measure that matters. It is also very important how the learned models were obtained, i.e., how much effort the tutor had to invest in order to teach the robot. Measuring the tutoring cost in such a mixed-initiative learning framework is quite a challenging problem; in this experiment we resorted to the following simple criterion: if the tutor had to provide the description of an object, it provided 3 bits of information (3 bits encode 8 classes of colours), while a polar answer was evaluated as a 1 bit cost. We therefore evaluated the different learning methods by comparing their recognition rate with respect to the cumulative tutoring costs. The evolution of the results over time is shown in Fig. 8.

In all experiments, we used 624 images for learning and the rest for testing. Each class was initialized by five labelled images. In the NS approach the first 60 samples were learnt in a tutor-driven mode to build initial models that were reliable enough to dictate the sequence of training images. The recognition performance was tested after every update on all test images.

All learning strategies reached the final recognition rate of 96%. This result is higher then the recognition rate presented in the system evaluation experiment, mainly because we used different parameter settings in the feature selection algorithm. For reference, we also trained the standard state-of-the-art SVM classifier using the RBF kernel. It produced inferior results, since it only reached 92% recognition rate. Therefore, in terms of the final recognition performance, all learning strategies were very successful.

The learning strategy TA was the most successful in terms of reaching top performance with minimal information provided. The strategies TD_{rnd}, TD_{seq} and NS were equal in amount of information provided by the tutor, but there is a striking difference in the learning rate. We can see that the order in which the images were presented, played a very important role. When the images were presented in sequential order ($TDseq$), the learning progress was very slow (since most of the training samples for some of the colours were presented towards the end of the learning sequence), while learning with the random sequence (TD_{rnd}) lead to significantly better performance. The tutor, would therefore have to pay a lot of attention to which of the objects to present. The NS approach achieved very similar results to the best TD approach; in this case, however, the sequence of learning was dictated by the system, which would relieve the tutor. We can expect that by combining these learning strategies we could achieve even better results.

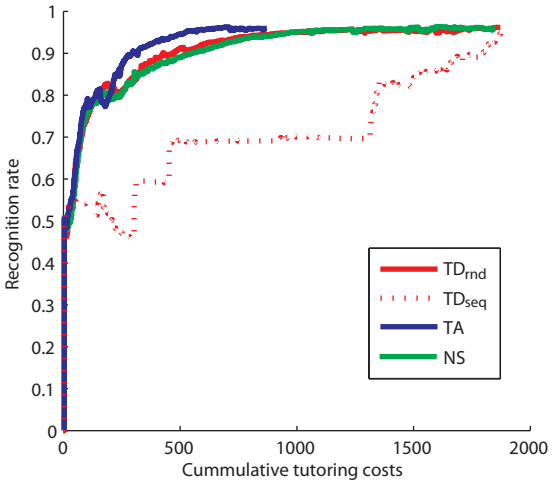

Fig. 8. Evaluation of different learning strategies.

V. 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. Due to lack of space we presented the capabilities of the developed system very briefly. We 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 drive – to extend its knowledge. We focused on three different types of learning mechanisms that are supported by the system. We demonstrated these principles in the case of learning about object colours and basic shapes.

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 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. Currently, we are making use of the robot's mobile platform and the pan-tilt unit to enable the robot to move and look around. This will increase the possibilities of interaction with the environment and enable the robot to acquire novel information in a 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 system, therefore, forms a firm basis for further development. 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 and A. P. Pentland, "Learning words from sights and sounds: a computational model," *Cognitive Science*, vol. 26, no. 1, pp. 113–146, 2002.

[2] L. Steels and F. Kaplan, "AIBO's first words, the social learning of language and meaning," *Evolution of Communication*, vol. 4, no. 1, pp. 3–32, 2000.

[3] C. Bauckhage, G. 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*, 2001, pp. 320–325.

[4] 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*, Daejeon, South Korea, Dec. 2008, pp. 553–560.

[5] A. L. Thomaz and C. Breazeal, "Experiments in socially guided exploration: lessons learned in building robots that learn with and without human teachers," *Connection Science*, vol. 20, no. 2 3, pp. 91–110, June 2008.

[6] S. Kirstein, A. Denecke, S. Hasler, H. Wersing, H.-M. Gross, and E. Körner, "A vision architecture for unconstrained and incremental learning of multiple categories," *Memetic Computing*, vol. 1, pp. 291–304, 2009.

[7] 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*, June 2009, pp. 1–6.

[8] K. Zhou, M. Zillich, M. Vincze, A. Vrečko, and D. Skočaj, "Multi-model fitting using particle swarm optimization for 3D perception in robot vision," in *IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 2010.

[9] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1222–1239, 2001.

[10] M. Kristan and A. Leonardis, "Online discriminative kernel density estimation," in *International Conference on Pattern Recognition*, Istanbul, Turkey, 23-26 August 2010, pp. 581–584.

[11] 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.

[12] M. Richardson and P. Domingos, "Markov logic networks," *Machine Learning*, vol. 62, no. 1-2, pp. 107–136, 2006.

[13] P. Lison, C. Ehrler, and G.-J. Kruijff, "Belief modelling for situation awareness in human-robot interaction," in *RO-MAN, 2010 IEEE*, 2010, pp. 138 –143.

[14] R. H. Thomason, M. Stone, and D. DeVault, "Enlightened update: A computational architecture for presupposition and other pragmatic phenomena," in *Presupposition Accommodation*, D. Byron, C. Roberts, and S. Schwenter, Eds. Ohio State Pragmatics Initiative, 2006.

[15] G.-J. Kruijff, M. Janíček, and P. Lison, "Continual processing of situated dialogue in human-robot collaborative activities," in *RO-MAN, 2010 IEEE*, 2010, pp. 594 –599.

[16] N. Hawes, "A survey of motivation frameworks for intelligent systems," *Artificial Intelligence*, vol. 175, no. 5-6, pp. 1020–1036, 2011.

[17] N. Hawes, M. Hanheide, J. Hargreaves, B. Page, H. Zender, and P. Jensfelt, "Home alone: Autonomous extension and correction of spatial representations," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA '11)*, May 2011.

[18] M. Helmert, "The fast downward planning system," *Journal of Artificial Intelligence Research*, vol. 26, pp. 191–246, 2006.

[19] M. Brenner and B. Nebel, "Continual planning and acting in dynamic multiagent environments," *Journal of Autonomous Agents and Multi-agent Systems*, vol. 19, no. 3, pp. 297–331, 2009.

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