A basic cognitive system for interactive continuous learning of visual concepts

Danijel Skočaj, Miroslav Janíček, Matej Kristan, Geert-Jan M. Kruijff,
Aleš Leonardis, Pierre Lison, Alen Vrečko, and Michael Zillich

***Abstract*— Interactive continuous learning is an important characteristic of a cognitive agent that is supposed to operate and evolve in an everchanging environment. In this paper we present representations and mechanisms that are necessary for continuous learning of visual concepts in dialogue with a tutor. We present an approach for modelling beliefs stemming from multiple modalities and we show how these beliefs are created by processing visual and linguistic information and how they are used for learning. We also present a system that exploits these representations and mechanisms, and demonstrate these principles in the case of learning about object colours and basic shapes in dialogue with the tutor.**

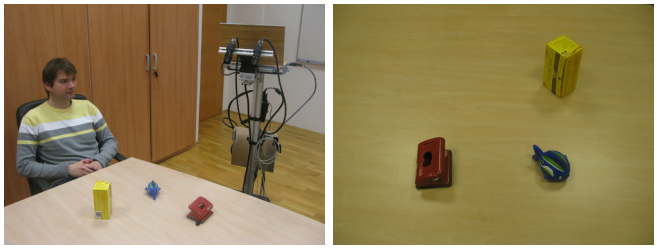
I. Introduction

An important characteristic of a cognitive system is the ability to continuously acquire new knowledge. Communication with a human tutor should significantly facilitate such incremental learning processes. In this paper we focus on representations and mechanisms that enable such interactive learning and present a system that was designed to acquire visual concepts through interaction with a human.

Such systems typically have several sources of information, vision and language being the most prominent ones. Based on the processed modal information corresponding beliefs are created that represent the robot's interpretation of the perceived environment. These beliefs rely on the particular representations of the perceived information in multiple modalities. These representations along with the cross-modal learning enable the robot to, based on interaction with the environment and people, extend its current knowledge by learning about the relationships between symbols and features that arise from the interpretation of different modalities. One modality may exploit information from another to update its current representations, or several modalities may be used together to form representations of a certain concept. We focus here on the former type of interaction between modalities and present the representations that are used for continuous learning of basic visual concepts in a dialogue with a human.

We demonstrate this approach on the robot George, which is engaged in a dialogue with the human tutor. Fig. 1 depicts a typical setup and the scene observed by the robot[1]. The main goal is to teach the robot about object properties (colours and two basic shapes). George has built-in abilities for visual processing and communication with a human, as well as learning abilities, however it does not have any model of object properties given in advance and therefore has to continuously build them. The tutor can teach the robot about object properties (e.g., 'H: This is a red thing.'), or the robot can try to learn autonomously or ask the tutor for help when necessary (e.g., 'G: Is the elongated thing red?'). Our aim is that the learning process is efficient in terms of learning progress, is not overly taxing with respect to tutor supervision and is performed in a natural, user friendly way.

In this paper we present the methodologies that enable such learning. First we present an approach for modelling beliefs stemming from multiple modalities in §II. We then show how these beliefs are used in dialogue processing in §III, followed by the description of representations and the learning process in vision in §IV. In §V we describe the system we have developed and in §VI we present an example of the scenario and the processing flow. We conclude the paper with a discussion and some concluding remarks.



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

Fig. 1. Continuous interactive learning of visual properties.

II. Modelling beliefs

High-level cognitive capabilities like dialogue operate on high level (i.e. abstract) representations that collect information from multiple modalities. Here we present an approach that addresses (1) how these high-level representations can be reliably generated from low-level sensory data, and (2) how information arising from different modalities can be efficiently fused into unified multi-modal structures.

The approach is based on a Bayesian framework, using insights from multi-modal information fusion [1], [2]. We

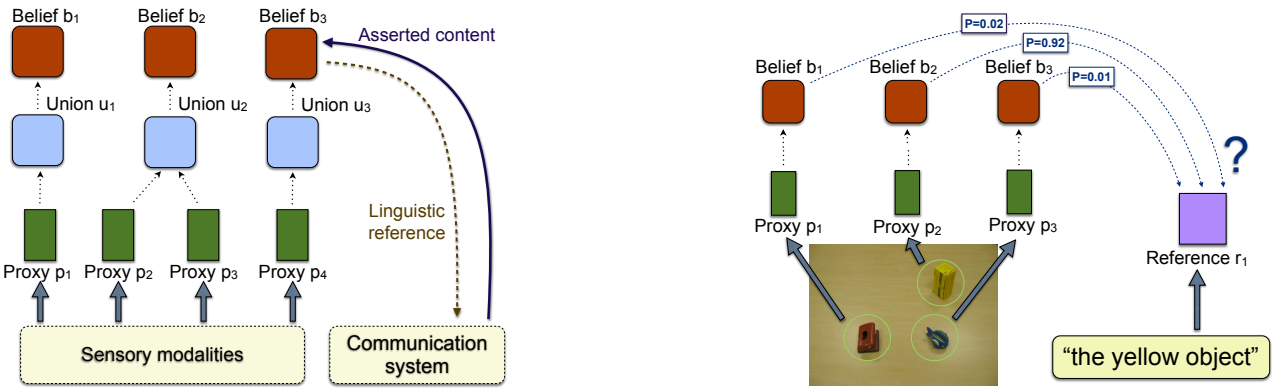
D. Skočaj, M. Kristan, A. Leonardis, and A. Vrečko are with University of Ljubljana, Slovenia, {danijel.skocaj, matej.kristan, ales.leonardis, alen.vrecko}@fri.uni-lj.si

M. Janíček, G.-J. M. Kruijff, and P. Lison are with DFKI, Saarbrücken, Germany, {miroslav.janicek, gj, plison}@dfki.de

M. Zillich is with Vienna University of Technology, Austria, zillich@acin.tuwien.ac.at

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

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



(a) Construction of beliefs.

(b) Reference resolution for the expression "the yellow object".

Fig. 2. Multi-modal information binding: belief construction (left) and application in a reference resolution task (right).

have implemented it as a specific subsystem called the *binder* [3]. The binder is linked to all other subsystems. It serves as a central hub for gathering information about entities currently perceived in the environment. The information on the binder is inherently probabilistic, so we can deal with varying levels of noise and uncertainty.

Based on the available information, the binder seeks to fuse the perceptual inputs arising from the various subsystems, by checking whether their respective features correlate with each other. The probability of these correlations are encoded in a Bayesian network. This Bayesian network can, for example, express a high compatibility between the haptic feature "shape: cylindrical" and the visual feature "object: mug" (since most mugs are cylindrical), but a very low compatibility between "shape: cylindrical" and "object: ball".

We call the resulting (amodal) information structure a *belief*. The task of the binder is to decide which perceptual inputs belong to the same real-world entity, and should therefore be unified into a belief. The outcome of this process is a joint probability distribution over possible beliefs. These beliefs integrate the information included in the perceptual inputs in a compact representation. They can therefore be directly used by the deliberative processes for planning, reasoning and learning.

In addition to the beliefs, there are two other central data structures manipulated by the binder, proxies and unions (see also Fig. 2(a)). A *proxy* is a uni-modal representation of a given entity in the environment. Proxies are inserted onto the binder by the various subarchitectures. They are defined as a multivariate probabilistic distribution over a set of features (discrete or continuous). A *union* is multi-modal representation of an entity, constructed by merging one or more proxies. Like proxies, unions are represented as a multivariate probabilistic distribution over possible features. They are essentially a transitional layer between proxies and beliefs.

A *belief* is an amodal representation of an entity in the environment. They are typically an abstraction over unions, expressed in an amodal format. A belief encodes additional information related to the specific situation and perspective in which the belief was formed. This includes its *spatio-temporal frame* (when and where and how an observation was made), its *epistemic status* (for which agents the belief holds, or is attributed), and a *saliency value* (a real-valued measure of the prominency of the entity [4]). Beliefs are indexed via a unique identifier, which allows us to keep track of the whole development history of a particular belief. Beliefs can also be connected with each other using relational structures of arbitrary complexity.

To create beliefs, the binder decides for each pair of proxies arising from distinct subsystems, whether they should be bound into a single union, or fork into two separate unions. The decision algorithm uses a technique from probabilistic data fusion, called the *Independent Likelihood Pool* (ILP) [5]. Using the ILP, we compute the likelihood of every possible binding of proxies, and use this estimate as a basis for constructing the beliefs. The multivariate probability distribution contained in the belief is a linear function of the feature distributions included in the proxies and the correlations between these. A Bayesian network encodes all possible feature correlations as conditional dependencies. The (normalised) product of these correlations over the complete feature set provides a useful estimate of the "internal consistency" of the constructed belief.

The beliefs, being high-level symbolic representations available for the whole cognitive architecture, provide a unified model of the environment which can be efficiently used when interacting with the human user.

III. Situated Dialogue

Situated dialogue provides one means for a robot to gain more information about the environment. A robot can discuss what it sees, and understands, with a human. Or it can ask about what it is unclear about, or would like to know more about.

That makes this kind of dialogue part of a larger activity. The human and the robot are working together. They interact to instruct, and to learn more. For that, they need to build up a common ground in understanding each other and the world.

Here we briefly discuss an approach that models dialogue as a collaborative activity. It models what is being said, and

why. It enables the robot to understand why it was told something, and what it needs to do with the information.

The approach is based on previous work by Stone & Thomason [6] (S&T). In their model, an agent uses abductive inference to construct an explanation of the possible intention behind a communicative act. This intention directs how an agent's belief models need to be updated, and what needs to be paid attention to next. This kind of inference is performed both for comprehension, and for production.

The problem with S&T is that they rely on a symmetry in communication: "What I say is how you understand it." This is untenable in human-robot interaction, particularly in a setting where a robot is learning about the world. Therefore, we have adapted and extended their approach to deal with (a) the asymmetry between what has been observed fact, and what has been asserted, and (b) clarification mechanisms, to overcome breakdowns in understanding.

Algorithm 1 Continual collaborative acting

$\Sigma^\pi = \emptyset$
loop {
Perception
  $e \leftarrow$ SENSE()
  $\langle c', i, \Pi \rangle \leftarrow$ UNDERSTAND($r, Z(c) \oplus \Sigma^\pi, e$)
  $c \leftarrow$ VERIFIABLE-UPDATE(c', i, Π)
Determination and Deliberation
  $c' \leftarrow$ ACT-TACITLY(p, c)
  $m \leftarrow$ SELECT(p, c')
  $\langle i, \Pi \rangle \leftarrow$ GENERATE($r, c', m, Z(c) \oplus \Sigma^\pi$)
Action
  ACT-PUBLICLY($a(i)$)
  $c \leftarrow$ VERIFIABLE-UPDATE(c', i, Π)
}

Algorithm 1 presents the core of the resulting model, based on S&T. In *perception*, the agent senses an event e. It tries to understand it in terms of an intention i that results in an update of the belief model from context c to c', given the communicative resources r, possible results $Z(c)$ to use them in context c, and whatever issues are still open to be resolved Σ^π. Given the inferred intention i and potential update c' the agent then tries to carry out this update, as a *verifiable update*. To model this, we use a logical framework of multi-agent beliefs (cf. §II) that includes a notion of *assertion* [7]. An assertion is a proposition that still needs to be verified. This verification can take various forms. In George, we check whether a new piece of information can be used to consistently update a belief model (consistency), or to extend a modal model (learning) or weaken it (unlearning). Any assertion still in need of verification ends up on Σ^π.

In *deliberation*, a tacit action based on some private information p is performed by the agent. In order to make the effects c' public, a public action m is selected and performed as a realisation $a(i)$ of the generated intention to act i.

An important aspect of linking dialogue with grounded beliefs is *reference resolution*: how to connect linguistic expressions such as "this box" or "the ball on the floor" to the corresponding beliefs about entities in the environment. The binder performs reference resolution using the same core mechanisms as used for binding. A Bayesian network specifies the correlations between the linguistic constraints of the referring expressions and the belief features (particularly, the entity saliency and associated categorical knowledge). Resolution yields a probability distribution over alternative referents (see Fig. 2(b) for an example). Abductive inference then determines which resolution hypothesis to use, in the context of establishing the best explanation. This is folded together with any new information an utterance might provide, to yield an update of the robot's current beliefs.

For example, consider an utterance like "This is yellow." First, the expression "this" must be resolved to a particular, proximal entity in the environment. Resolution is performed on the basis of the saliency measures. Second, the utterance also provides new information about the entity, namely that it is yellow. The robot's beliefs get updated with this asserted information. Dialogue processing does this by selecting the belief about the referred-to entity, then incorporating the new information. Indirectly, this acts as a trigger for learning.

In George, the dynamics of assertions on Σ^π provide the main drive for how learning and dialogue interact. The vision subarchitecture can pose *clarification requests* to the dialogue system. These requests are interpreted as tacit actions (Algorithm 1), pushing an assertion onto Σ^π. This assertion may be a polar or an open statement. Then similarly to resolving any breakdown in understanding the user, the robot can decide to generate a clarification subdialogue. This dialogue continues until the (original) assertion has been verified, i.e. a proper answer has been found [8].

IV. Learning visual concepts

In the two previous sections we discussed how the modal information gathered from individual modalities is fused into unified multi-modal structures and how they are used in situated dialogue. In this section we will describe how the modal information is captured and modelled in the visual subarchitecture; how these models are initiated and how they are being continuously updated and how they can be queried to provide the abstracted information for higher-level cognitive processing.

To efficiently store and generalise the observed 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 highdimensional features (e.g., multiple 1D features relating to shape, texture, colour and intensity of the observed object) and the online Kernel Density Estimator (oKDE) [9] is used to estimate the pdf in this high-dimensional feature space. The oKDE estimates the probability density functions by a mixture of Gaussians, is able to adapt using only a single data-point at a time, automatically adjusts its complexity and does not assume specific requirements on the target distribution. A particularly important feature of the oKDE is that it allows adaptation

from the positive examples (learning) as well as negative examples (unlearning) [10].

However, concepts such as *colour red* reside only within a lower dimensional subspace spanned only by features that relate to colour (and not texture or shape). Therefore, during the learning, this subspace has to be identified to provide the best performance. This is achieved by determining the optimal subspace for a set of mutually exclusive concepts (e.g., all colours, or all shapes). We assume that this corresponds to the subspace which minimises the overlap of the corresponding distributions. The overlap between the distributions is measured using the multivariate Hellinger distance [9]. An example of the learnt models is shown in Fig. 3.

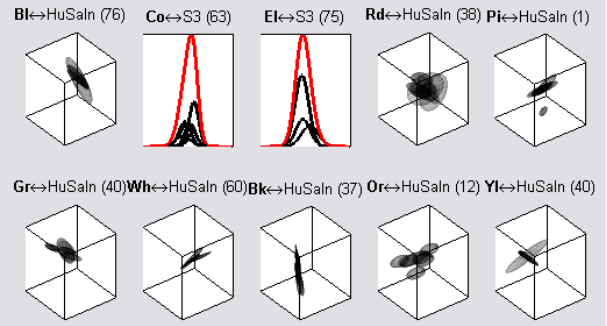


Fig. 3. Example of the models estimated using the oKDE and the feature selection algorithm. Note that some concepts are modelled by 3D distributions (e.g., "blue" which is denoted by "Bl"), while others (e.g., "compact" which is denoted by "Co") is modelled by 1D distributions.

Therefore, during online operation, a multivariate generative model is continually maintained for each of the visual concepts and for mutually exclusive sets of concepts the feature subspace is continually being determined. This feature subspace is then used to construct a Bayesian classifier for a set of mutually exclusive concepts, which can be used for recognition of individual object properties.

However, since the system is operating in an online manner, the closed-world assumption cannot be assumed; at every step the system should also take into account the probability that it has encountered a concept that has not been observed before. Therefore, when constructing the Bayesian classifier, an "unknown model" has also to be considered besides the learned models. It should account for a poor classification when none of the learnt models supports the current observation strongly enough. We assume that the probability of this event is uniformly distributed over the feature space. The a priori probability of the "unknown model" is assumed to be non-stationary and decreases with increasing numbers of observations; the more training samples the system observes, the smaller is the probability that it will encounter something novel.

Having built such a knowledge model and Bayesian classifier, the recognition is done by inspecting a posteriori probability (AP) of individual concepts and unknown model; in fact the AP distribution over the individual concepts is packed in a vision proxy, which is sent to the binder and serves as a basis for forming a belief about the observed object as described in §II (see also Fig. 2(b)).

Furthermore, such a knowledge model is also appropriate for detecting incompleteness in knowledge. It can be discovered through inspection of the AP distribution. In particular, we can distinguish two general cases. (1) In the first case the observation can be best explained by the unknown model, which indicates a gap in the knowledge; the observation should most probably be modeled with a model that has not yet been learned. A clarification request is issued that results in an open question (e.g., 'Which colour is this?'). (2) In the second case the AP of the model that best explains the observation is low, which indicates that the classification is very uncertain and that the current model cannot provide a reliable result. A clarification request is issued that results in a polar question (e.g., 'Is this red?'). In both cases, after the tutor provides the answer, the system gets the additional information, which allows it to improve the model by learning or unlearning.

V. System Architecture

We have implemented the representations and mechanisms described in the previous sections in the robot George. In this section we describe the system architecture and the individual components that are involved.

For implementation of the robot we employ a specific architecture schema, which we call CAS (CoSy Architecture Schema) [11]. The schema is essentially a distributed working memory model, where representations are linked within and across the working memories, and are updated asynchronously and in parallel. The system is therefore composed of several subarchitectures implementing different functionalities and communicating through their working memories. The George system is composed of three such subarchitectures: the *Binder SA*, the *Communications SA* and the *Visual SA*, as depicted in Fig. 4. Here, the components of the visual subsystem could be further divided into three distinct layers: the quantitative layer, the qualitative layer and the mediative layer.

In the previous subsections we assumed that the modal information is adequately captured and processed. Here we briefly describe how the relevant visual information is detected, extracted and converted in the form that is suitable for processing in the higher level processes. This is the task of *the quantitative layer* in the Visual SA. The quantitative layer processes the visual scene as a whole and implements one or more *bottom-up* visual attention mechanisms. A bottom-up attention mechanism tries to identify regions in the scene that might be interesting for further visual processing. George currently has one such mechanism, which uses the stereo 3D point cloud provided by *stereo reconstruction component* to extract the dominant planes and the things sticking out from those planes. Those sticking-out parts form spherical 3D spaces of interest (SOIs). The *SOI Analyzer* component validates the SOIs and, if deemed interesting (considering SOI persistence, stability, size, etc.), upgrades them to *proto-objects* adding information that is needed for the qualitative

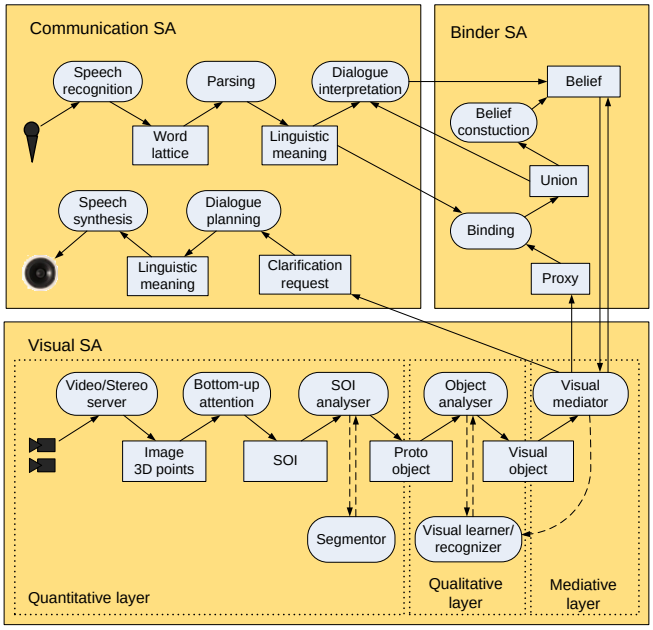


Fig. 4. Architecture of the George system.

processing, e. g. the object segmentation mask (the proto-object is segmented by the Graph cut algorithm [12] using the 3D and colour information provided by the stereo reconstruction).

The qualitative layer implements the main functionalities for recognition and learning of visual concepts that were described in §IV. This layer processes each interesting scene part (object) individually, focusing on qualitative properties. After the extraction of the visual attributes (in the *Visual Learner-recognizer*), like color and shape, the *Object Analyzer* upgrades the proto-objects to *visual objects*. Visual objects encapsulate all the information available within the Visual SA and are the final modal representations of the perceived entities in the scene. Also, the learning of visual attributes is performed in this layer.

The main purpose of *the mediative layer* is to exchange information about the perceived entities with other modalities. This is not done directly, but via the specialised a-modal subarchitecture Binder SA, that actually creates and maintains beliefs as described in §II. The *Visual Mediator component* adapts and forwards the modal information about objects to the binder (each visual object is represented by a dedicated proxy in the binder). The component also monitors beliefs for possible learning opportunities, which result in modal learning actions. Another important functionality of the mediator is to formulate and forward clarification motivations in the case of missing or ambiguous modal information. Currently, these motivations are directly intercepted by the Communication SA.

Given a clarification request, the *Communication SA* formulates a dialogue goal given the information the system needs to know and how that can be related to the current dialogue and belief-context. Dialogue planning turns this goal into a meaning representation that expresses the request in context. This is then subsequently synthesised, typically as a question about a certain object property. When it comes to understanding, the Communication SA analyses an incoming audio signal and creates a set of possible word sequences for it. This is represented as a word lattice, with probabilities indicating the likelihood that a certain word was heard, in a particular sequence. The word lattice is then subsequently parsed, and from the space of possible linguistic meaning representations for the utterance, the contextually most appropriate one is chosen [13]. Finally, dialogue interpretation takes the selected linguistic meaning. This meaning is then interpreted against a belief model, to understand the intention behind the utterance. We model this is an operation on how the system's belief model is intended to be updated with the information provided. In §VI below we provide more detail, given an example.

VI. Example scenario

A. Scenario setup

The robot operates in a table-top scenario, which involves a robot and a human tutor (see Fig. 1(a)). The robot is asked to recognise and describe the objects in the scene (in terms of their properties like colour and shape). The scene contains a single object or several objects, with limited occlusion. The human positions new objects on the table and removes the objects from the table while being involved in a dialogue with the robot. In the beginning the robot does not have any representation of object properties, therefore it fails to recognise the objects and has to learn. To begin with, the tutor guides the learning process and teaches the robot about the objects. After a while, the robot takes the initiative and tries to detect its own ignorance and to learn autonomously, or asks the tutor for assistance when necessary. The tutor can supervise the learning process and correct the robot when necessary; the robot is able to correct erroneously learned representations. The robot establishes transparency and verbalises its knowledge and knowledge gaps. In a dialogue with the tutor, the robot keeps extending and improving the knowledge. The tutor can also ask questions about the scene, and the robot is able to answer (and keeps giving better and better answers). At the end, the representations are rich enough for the robot to accomplish the task, that is, to correctly describe the initial scene.

B. Example script

Two main types of learning are present in the George scenario, which differ on where the motivation for a learning update comes from. In tutor driven learning the learning process is initiated by the human teacher, while in tutor assisted learning, the learning step is triggered by the robot.

Tutor driven learning is suitable during the initial stages, when the robot has to be given information, which is used to reliably initiate (and extend) visual concepts. Consider a scene with a single object present:

H: Do you know what this is?
G: No.
H: This is a red object.

G: Let me see. OK.

Since in the beginning, George doesn't have any representation of visual concepts, he can't answer the question. After he gets the information, he can first initiate and later sequentially update the corresponding information.

After a number of such learning steps, the acquired models become more reliable and can be used to reference the objects. Therefore, there can be several objects in the scene, as in Fig. 1, and George can talk about them:

H: What colour is the elongated object?

G: It is yellow.

When the models are reliable enough, George can take the initiative and try to learn without being told to. In this curiosity-driven learning George can pose a question to the tutor, when he is able to detect the object in the scene, but he is not certain about his recognition. As described in §IV in such *tutor-assisted* learning there are two general cases of detection of uncertainty and knowledge gaps. If the robot cannot associate the detected object with any of the previously learned models, it considers this as a gap in its knowledge and asks the tutor to provide information:

R: Which colour is this object?

H: It is yellow.

R. OK.

The robot is now able to initialise 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 recognise 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 orange.

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 orange and makes these two representations more reliable.

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: Do you know what this is?

G: It is a blue object.

H: What shape is the green object?

G: It is elongated.

C. Processing flow

Here we describe the processing flow for one illustrative example. We describe in more detail what happens after the human places several objects in the scene (see Fig. 1) and refers to the only elongated object in the scene (the yellow tea box) by asserting "*H: The elongated object is yellow.*".

In the Visual SA the tea box is represented by a *SOI* on the quantitative layer, a *proto-object* on the qualitative layer and a *visual object* on the mediative layer. Let us assume that the *Visual Learner-recognizer* has recognised the object as being of elongated shape, but has completely failed to recognise the colour. In the binder this results in a one-proxy union with the binding features giving the highest probability to the elongated shape, while the colour is considered to be unknown. This union is referenced by the single robot's private belief in the belief model (Fig. 5, step 1).

The tutor's utterance 'The elongated object is yellow' is processed by the Communication SA. Speech recognition turns the audio signal into a set of possible sequences of words, represented as a word lattice. The Communication SA parses this word lattice incrementally, constructing a representation of the utterance's most likely linguistic meaning in context [13]. We represent this meaning as a logical form, an ontologically richly sorted relational structure. Given this structure, the Communication SA establishes which meaningful parts might be referring to objects in the visual context. For each such part, the binder then computes possible matches with unions present in the binding memory, using phantom proxies (Fig. 5, step 2). These matches form a set of reference hypotheses. The actual reference resolution then takes place when we perform dialogue interpretation. In this process, we use weighted abductive inference to establish the intention behind the utterance – why something was said, and how the provided information is to be used. The proof with the lowest cost is chosen as the most likely intention. Reference resolution is done in this larger context of establishing the "best explanation." Abduction opts for that referential hypothesis which leads to the overall best proof. The resulting proof provides us with an intention, and a belief attributed to the tutor is constructed from the meaning of the utterance. In our example, this attributed belief restricts the shape to elongated, asserts the colour to be yellow and references the union that includes the visual proxy representing the yellow tea box.

In the Visual SA, the mediator intercepts the event of adding the attributed belief. The colour assertion and the absence of the colour restriction in the robot's belief is deemed as a learning opportunity (the mediator knows that both beliefs reference the same binding union, hence the same object). The mediator translates the asserted colour information to an equivalent modal colour label and compiles a learning task. The learner-recognizer uses the label and the lower level visual features of the tea box to update its yellow colour model. After the learning task is complete, the mediator verifies the attributed belief, which changes its epistemic status to shared (Fig. 5, step 3). The learning action re-triggers the recognition. If the updated yellow colour model is good enough, the colour information in the binder and belief model is updated (Fig. 5, step 4).

A similar process also takes place in tutor assisted learning when the robot initiates the learning process, based on an unreliable recognition, e.g., by asking "*R: Is this red?*". In this case, the need for assistance reflects in a robot's private belief that contains the assertion about the red colour and references the union representing the object. Based on this belief, the Communication SA synthesises the above ques-

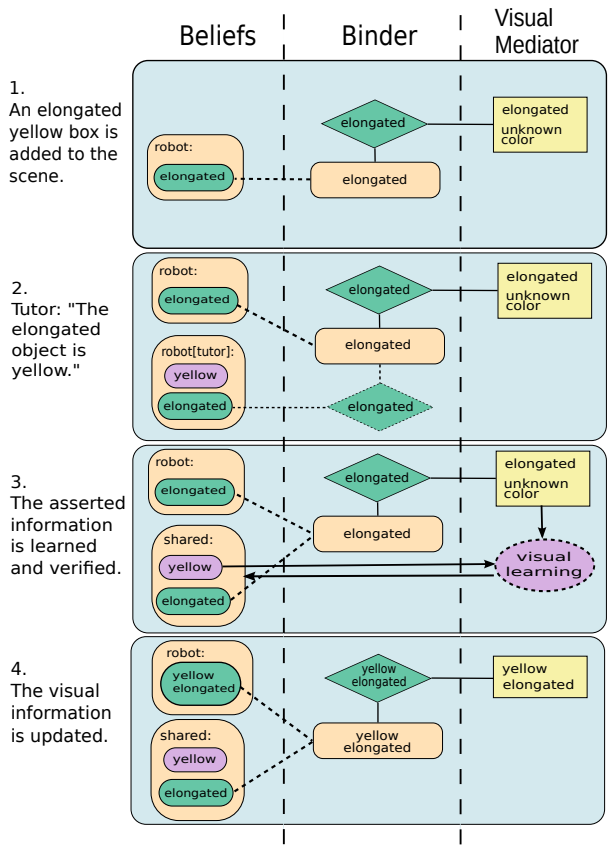


Fig. 5. Example of processing flow in the binder. The green colour represents restrictive information, while the violet colour denotes assertive information. Only the beliefs and other data structures pertaining to the yellow tea box are shown.

tion. When the robot receives a positive answer, it updates the representation of red, using a very similar mechanism as in the case of tutor driven learning.

VII. Conclusion

In this paper we presented representations and mechanisms that are necessary for continuous learning of visual concepts in dialogue with a tutor. An approach for modelling beliefs stemming from multiple modalities was presented and it was shown how these beliefs are created by processing visual and linguistic information and how they are used for learning. We also presented a system that exploits these representations and mechanisms and demonstrated these principles in the case of learning about object colours and basic shapes in a dialogue with the tutor.

We have made several contributions at the level of individual components (modelling beliefs, dialogue processing, incremental learning), as well as at the system level (by integrating the individual components in a coherent multi-modal distributed asynchronous system). Such an integrated robotic implementation enables system-wide research with all its benefits (information provided by other components), as well its problems and challenges (that do not occur in simulated or isolated environments). We are, therefore, now able to directly investigate the relations between the individual components and analyse the performance of the robot at the sub-system and system level. This will allow us to set new requirements for individual components and to adapt the components, which will result in a more advanced and robust system.

The main goal was to set up a framework that would allow the system to process, to fuse, and to use the information from different modalities in a consistent and scalable way on different levels of abstraction involving different kinds of representations. This framework has been implemented in the robot George, which is still limited in several respects; it operates in a constrained environment, the set of visual concepts that are being learned is relatively small, and the mixed initiative dialogue is not yet matured. We have been addressing these issues and the robot will gradually become more and more competent. Furthermore, we also plan to integrate other functionalities that have been under development, like motivation and manipulation.

The presented system already exhibits several properties that we would expect from a cognitive robot that is supposed to learn in interaction with a human. As such, it 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 everchanging world by capturing and processing cross-modal information in an interaction with the environment and other cognitive agents.

References

[1] D. Roy, "Semiotic schemas: A framework for grounding language in action and perception," *Artificial Intelligence*, vol. 167, no. 1-2, pp. 170–205, 2005.

[2] R. Engel and N. Pfleger, "Modality fusion," in *SmartKom: Foundations of Multimodal Dialogue Systems*, W. Wahlster, Ed. Berlin: Springer, 2006, pp. 223–235.

[3] H. Jacobsson, N. Hawes, G.-J. Kruijff, and J. Wyatt, "Crossmodal content binding in information-processing architectures," in *Proc. of the 3rd International Conference on Human-Robot Interaction*, 2008.

[4] J. Kelleher, "Integrating visual and linguistic salience for reference resolution," in *Proceedings of the 16th Irish conference on Artificial Intelligence and Cognitive Science (AICS-05)*, N. Creaney, Ed., 2005.

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

[6] R. Thomason, M. Stone, and D. DeVault, "Enlightened update: A computational architecture for presupposition and other pragmatic phenomena," in *Presupposition Accommodation*, to appear.

[7] M. Brenner and B. Nebel, "Continual planning and acting in dynamic multiagent environments," *Journal of Autonomous Agents and Multi-agent Systems*, 2008.

[8] G. Kruijff and M. Brenner, "Phrasing questions," in *Proceedings of the AAAI 2009 Spring Symposium on Agents That Learn From Humans*, 2009.

[9] M. Kristan and A. Leonardis, "Multivariate online kernel density estimation," in *Computer Vision Winter Workshop*, 2010, pp. 77–86.

[10] M. Kristan, D. Skočaj, and A. Leonardis, "Online kernel density estimation for interactive learning," *Image and Vision Computing*, 2009.

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

[12] Y. Boykov, O. Veksler, and R. Zabih, "Efficient approximate energy minimization via graph cuts," *PAMI*, vol. 20, no. 12, pp. 1222 – 1239, 2001.

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