Belief Modelling for Situation Awareness in Human-Robot Interaction

Pierre Lison, Carsten Ehrler and Geert-Jan M. Kruijff

***Abstract*— To interact naturally with humans, robots needs to be aware of their own surroundings. This awareness is usually encoded in some implicit or explicit representation of the situated context. In this paper, we present a new framework for constructing rich belief models of the robot's environment. Key to our approach is the use of *Markov Logic* as a unified framework for inference over these beliefs. Markov Logic is a combination of first-order logic and probabilistic graphical models. Its expressive power allows us to capture both the rich relational structure of the environment and the uncertainty arising from the noise and incompleteness of low-level sensory data. The constructed belief models evolve dynamically over time and incorporate various contextual information such as spatio-temporal framing, multi-agent epistemic status, and saliency measures. Beliefs can also be referenced and extended "top-down" via linguistic communication. The approach is being integrated into a cognitive architecture for mobile robots interacting with humans using spoken dialogue.**

I. Introduction

The situated context plays a central role in human-robot interaction (HRI). To be able to interact naturally with humans, robots needs to be aware of their own environment. This situation awareness is generally expressed in some sort of *belief models* in which various aspects of the external reality are encoded. Such belief models provide an explicit or implicit representation for the current state of the world, from the robot's viewpoint. They therefore serve as a representational backbone for a wide range of high-level cognitive capabilities related to reasoning, planning and learning in complex and dynamic environments. They are also essential for the robot to verbalise its own knowledge.

In speech-based HRI, critical tasks in dialogue understanding, management and production are directly dependent on such belief models. Examples are context-sensitive speech recognition [15], reference resolution and generation in small- [11] and large-scale space [24], spoken dialogue parsing [14] and interpretation [20], dialogue management [23], user-tailored response generation [22], and contextually appropriate intonation patterns [13]. Contextual knowledge is also a prerequisite for the dynamic adaptation of the robot's behaviour to different environments and interlocutors [3].

Belief models are usually expressed as high level symbolic representations merging and abstracting information over multiple modalities. For HRI, the incorporated knowledge might include (inter alia): entities in the visual scene, spatial structure, user profiles (intentional and attentional state, preferences), dialogue histories, and task models (what is to be done, which actions are available).

The construction of such belief models raises two important issues. The first question to address is how these high-level representations can be reliably abstracted from low-level sensory data [1], [18]. To be meaningful, most symbolic representations should be *grounded* in (subsymbolic) sensory inputs [19]. This is a difficult problem, partly because of the noise and uncertainty contained in sensory data (partial observability), and partly because the connection between low-level perception and high-level symbols is typically difficult to formalise in a general way [6].

The second issue relates to how information arising from different modalities and time points can be efficiently *merged* into unified multi-modal structures [12], and how these inputs can refine and constrain each other to yield improved estimations, over time. This is the well-known engineering problem of multi-target, multi-sensor data fusion [5].

Belief models are thus the final product of an iterative process of information *fusion*, *refinement* and *abstraction*. Typical HRI environments are challenging to model, being simultaneously *complex*, *multi-agent*, *dynamic* and *uncertain*. Four requirements can be formulated:

1) HRI environments are complex and reveal a large amount of internal structure (for instance, spatial relations between entities, or groupings of objects). The formal representations used to model them must therefore possess the expressive power to reflect this rich relational structure.
2) Interactive robots are made for multi-agent settings. Making sense of communicative acts requires the ability to distinguish between one's own knowledge (what I believe), knowledge attributed to others (what I think the others believe), and shared common ground knowledge (what we believe as a group).
3) Situated interactions are *dynamic* and evolve over time. The incorporation of spatio-temporal framing is thus necessary to go beyond the "here-and-now" and be capable of linking the present with (episodic) memories of the past and anticipation of future events.
4) And last but not least, due to the partial observability of most contextual features, it is crucial that belief models incorporate an explicit account of *uncertainties*.

Orthogonal to these "representational" requirements, crucial performance requirements must also be adressed. To keep up with a continuously changing environment, all operations on belief models (updates, queries, etc.) must be performed under soft real-time constraints.

This work was supported by the EU FP7 ICT Integrated Project "*CogX: cognitive systems that self-understand and self-extend*" (FP7-ICT- 215181).

Pierre Lison, Carsten Ehrler and Geert-Jan M. Kruijff are with the German Research Centre for Artificial Intelligence (DFKI GmbH), Language Technology Lab, Saarbrücken, Germany. {plison,carsten.ehrler,gj}@dfki.de

This paper presents ongoing work on a new approach to multi-modal situation awareness which attempts to address these requirements. Key to our approach is the use of a first-order probabilistic language, *Markov Logic* [17], as a unified representation formalism to perform various kind of inference over rich, multi-modal models of context. Markov Logic is a combination of first-order logic and probabilistic modelling. As such, it provides an elegant account of both the uncertainty and complexity of situated human-robot interactions. Our approach departs from previous work such as [9] or [18] by introducing a much richer modelling of multi-modal beliefs. Multivariate probability distributions over possible values are used to account for the partial observability of the data, while the first-order expressivity of Markov Logic allows us to consisely describe and reason over complex relational structures. As we shall see, these relational structures are annotated with various contextual information such as spatio-temporal framing (where and when is the entity assumed to exist), epistemic status (for which agents does this belief hold), and saliency (how prominent is the entity relative to others). Furthermore, performance requirements can be addressed with approximation algorithms for probabilistic inference optimised for Markov Logic [17], [16]. Such algorithms are crucial to provide an upper bound on the system latency and thus preserve its efficiency and tractability.

The rest of this paper is structured as follows. Section II provides a brief introduction to Markov Logic, the framework used for belief modelling. Section III details our approach in terms of architecture, representations, and processing operations. Section IV discusses further aspects of our approach. Section V concludes and provides directions for future work.

II. Background

Markov logic combines first-order logic and probabilistic graphical models in a unified representation [17]. A *Markov logic network* L is defined as a set of pairs (F_i, w_i), where F_i is a first-order formula and $w_i \in \mathbb{R}$ is the associated weight of that formula. A Markov logic network can be interpreted as a *template* for constructing Markov networks, which in turn can be used to perform probabilistic inference over the relational structure defined by the set of formulas F_i.

A. Markov Network

A Markov network G, also known as a *Markov random field*, is an undirected graphical model [10] for the joint probability distribution of a set of random variables $X = (X_1, \ldots, X_n) \in \mathcal{X}$. The network G contains a node for each random variable X_i. The joint probability of a Markov network is defined as such:

$$P(X = x) = \frac{1}{Z} \prod_k \phi_k(x_{\{k\}}) \tag{1}$$

where $\phi_k(x_{\{k\}})$ is a *potential function* mapping the state of a clique[1] k to a non-negative real value. Z is a normalization constant (known as *partition function*).

[1]In graph theory, a clique is a fully connected subgraph. That is, a subset of nodes where each node is connected with each other.

Alternatively, the potential function ϕ_k in (1) can be replaced by an exponentiated weighted sum over real-valued feature functions f_j:

$$P(X = x) = \frac{1}{Z} e^{(\sum_j w_j f_j(x))} \tag{2}$$

B. Ground Markov Network

Recall that a Markov logic network L is a set of pairs (F_i, w_i). If in addition to L we also specify a set of constants $C = \{c_1, c_2, ..., c_{|C|}\}$, one can generate a *ground Markov network* $M_{L,C}$ as follows:

1) For each possible predicate grounding over the set C, there is a binary node in $M_{L,C}$. The value of the node is true iff the ground predicate is true.
2) For every formula F_i, there is a feature f_j for each possible grounding of F_i over C. The value of the feature $f_i(x)$ is 1 if F_i is true given x and 0 otherwise. The weight of the feature corresponds to the weight w_i associated with F_i.

The graphical representation of $M_{L,C}$ contains a node for each ground predicate. Furthermore, each formula F_i defines a set of cliques j with feature f_j over the set of distinct predicates occurring in F_i. For further details see [17].

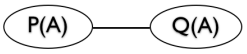


Fig. 1. Example (adapted from [17]) of a ground Markov Network $M_{L,C}$ given the Markov logic network $L = (\forall x.P(x) \vee Q(x), w)$ and $C = \{A\}$. It contains a single clique with feature f. The value of f is 1 for the three worlds $(P(A), Q(A)), (\neg P(A), Q(A)), (P(A), \neg Q(A))$. Following Eq. (3), the probability of each of these worlds is e^w/Z, where $Z = e^w + 1$. For the last world $(\neg P(A), \neg Q(A))$ the formula is false ($f = 0$) and its probability is $1/Z$ (thus tending to 0 as $w \to \infty$).

C. Inference

Once a Markov network $M_{L,C}$ is constructed, it can be exploited to perform conditional inference over the relational structure defined by L. Following (1), the joint probability distribution of a ground Markov network $M_{L,C}$ is given by

$$P(X = x) = \frac{1}{Z} \prod_i \phi_i(x_{\{k\}})^{n_i(x)} = \frac{1}{Z} e^{(\sum_i w_i n_i(x))} \tag{3}$$

The function $n_i(x)$ in (3) counts the number of true groundings of the formula F_i in $M_{L,C}$ given x. Due to the normalization term Z, exact inference is in general infeasible. However, efficient algorithms for probabilistic inference such as Markov Chain Monte Carlo (MCMC) can then be used to yield approximate solutions [16].

D. Learning

The weight w_i in a Markov logic network encode the "strength" of its associated formula F_i. In the limiting case, where $\lim_{w_i \to \infty}$, the probability of a world violating F_i has zero probability. For smaller values of the weight, worlds violating the formula will have a low, but non-zero probability. Weights can be learned on training samples using classical gradient-based techniques, or sampling.

III. Approach

We now describe our approach to belief modelling for situation awareness. We detail the architecture in which our system is integrated, the representations we used, and the processing components operating on them.

A. Architecture

Our approach is being developed as part of a distributed cognitive architecture for autonomous robots in open-ended environments [7]. The architecture has been applied to various scenarios such as visual learning and object manipulation in a tabletop scene [21] and exploration of indoor environments for human-augmented mapping [8].

Our approach to rich multi-modal belief modelling is implemented in a specific module called the "*binder*". The binder is directly connected to all subsystems in the architecture (i.e. vision, navigation, manipulation, etc.), and serves as a central hub for the information gathered about the environment. The core of the binder system is a shared *working memory* where beliefs are formed and refined based on incoming perceptual inputs. Fig. 2 illustrates the connection between the binder and the rest of the architecture.

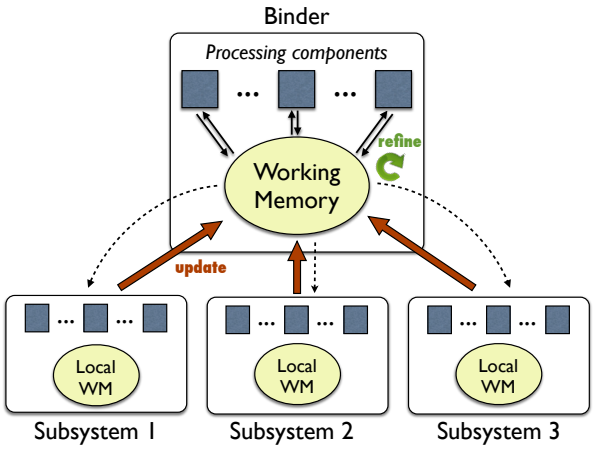


Fig. 2. Schema of the cognitive architecture in relation with the binder system and its working memory

B. Representation of beliefs

Each unit of information describing an entity[2] is expressed as a *probability distribution* over a space of alternative values. These values are formally expressed as propositional logical formulae. Such unit of information is called a **belief**.

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 information contained in the belief holds. Finally, beliefs are also given an *ontological category* used to sort the various belief types.

Formally, a belief is a tuple $\langle i, e, \sigma, c, \boldsymbol{\delta} \rangle$, where i is the belief identifier, e is an epistemic status, σ a spatio-temporal frame, c an ontological category, and $\boldsymbol{\delta}$ is the belief content itself. The content $\boldsymbol{\delta}$ is typically defined by a list of features. For each feature, we have a (continuous or discrete) distribution over alternative values. Fig. 3(a) provides a schematic illustration of a belief.

In addition, beliefs also contain bookkeeping information detailing the history of their formation. This is expressed as pointers to the belief ancestors (i.e. the beliefs which contributed to the emergence of this particular belief) and offspring (the ones which themselves emerged out of it).

The spatio-temporal frame σ defines a a probability distribution over the existence of the entity in a given temporal and spatial domain. The frame can for instance express that a particular visual object is thought to exist (with a given probability) in the world at a location l and in a temporal interval $[t_1, t_2]$.

The epistemic status e for an agent a can be either:

- *private*: denoted $\{a\}$, is a result of agent a's perception of the environment;
- *attributed*: denoted $\{a[b_1, ..., b_n]\}$, is a's conjecture about the mental states of other agents $b_1, ..., b_n$, usually resulting from communicative acts.
- *shared*: denoted $\{a_1, ..., a_m\}$, is information which is part of the common ground for the group[2].

As a brief illustration, assume a belief b_i defined as

$$\langle i, \{\text{robot}\}, \sigma_i, \mathsf{visualobject}, \boldsymbol{\delta}_i \rangle \tag{4}$$

where the spatio-temporal frame σ_i can be a normal distribution over 3D space combined with a temporal interval:

$$\sigma_i = (\mathcal{N}_3(\boldsymbol{\mu}, \boldsymbol{\Sigma}), [t_1, t_2]) \tag{5}$$

and with the content $\boldsymbol{\delta}_i$ being composed of two features:

$$\langle \textsc{Label} \rangle = \{(\mathsf{mug}, 0.7), (\mathsf{Unknown}, 0.3)\} \tag{6}$$

$$\langle \textsc{Colour} \rangle = \{(\mathsf{red}, 0.8), (\mathsf{orange}, 0.2)\} \tag{7}$$

Note that the probability distributions between features are by default assumed to be conditionally independent.

Feature values can be either discrete (as for categorical knowledge) or continuous (as for real-valued measures). A feature value can also be a pointer to another formula:

$$\langle \textsc{Location} \rangle \; k \tag{8}$$

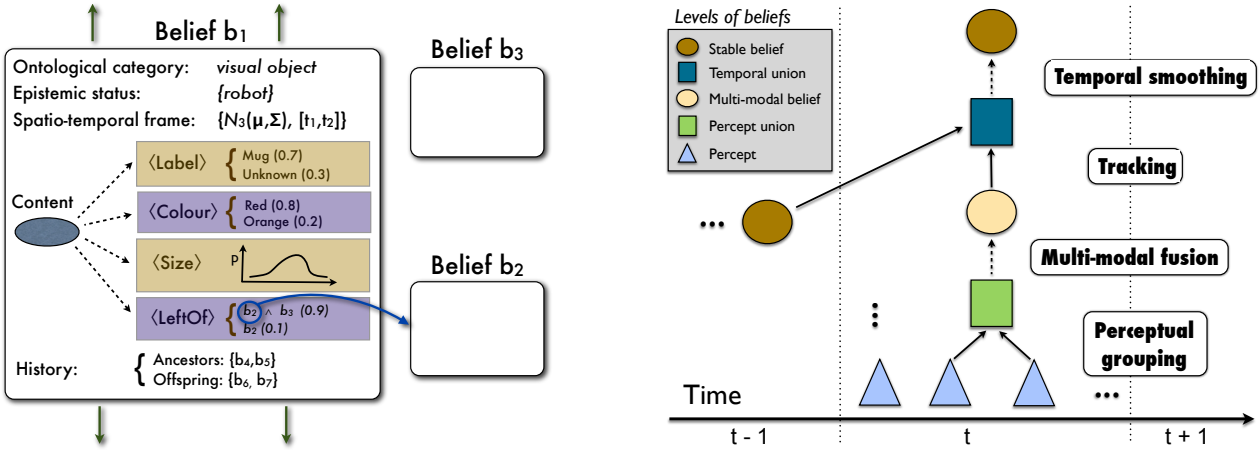
where k points to another belief. Such pointers are crucial to capture relational structures between entities.

Converting the probability distribution $\boldsymbol{\delta}$ into Markov Logic is relatively straightforward. Modal operators are translated into first-order predicates and nominals into contants. A (sub-)formula $\langle \textsc{Colour} \rangle$ blue with probability p_1 for a belief i is therefore expressed as:

$$w_1 \quad \mathtt{Colour}(\mathtt{I}_1, \mathtt{I}_2) \wedge \mathtt{Blue}(\mathtt{I}_2) \tag{9}$$

where the weight $w_1 = \log \dfrac{p_1}{1 - p_1}$.

[2]The term "entity" should be understood here in a very general sense. An entity can be an object, a place, a landmark, a person, etc.



(a) Schematic view of a belief representation.

(b) Bottom-up belief formation.

Fig. 3. Rich belief modelling for HRI: representations (left) and processing (right).

C. Levels of beliefs

The beliefs constructed and refined in the binder can be of different types. The number and nature of these types depend on the application domain. We discuss here four levels which are common for cognitive robotic architectures:

1) The lowest-level type of beliefs is the *percept*, which is a uni-modal representation of a given entity in the environment. Perceptual beliefs are inserted onto the binder by the various subsystems included in the architecture. The epistemic status of a percept is private per default, and the temporal frame is constrained to the present time-point.
2) If several percepts (from distinct modalities) are assumed to originate from the same entity, they can be grouped into a *percept union*. A percept union is just another belief, whose content is the combination of all the features from the included percepts.
3) The features of a percept union can be abstracted using multi-modal fusion and yield a *multi-modal belief*.
4) If the current multi-modal belief (which is constrained to the present spatio-temporal frame) is combined with beliefs encoded in past or future spatio-temporal frames, it forms a *temporal union*.
5) Finally, the temporal unions can be refined *over time* to improve the estimations, leading to a *stable belief*, which is both multi-modal and spans an extended temporal frame.

Since beliefs can point to each other, such models are able to capture relational structures of arbitrary complexity. Beliefs can also express past or future knowledge (i.e. memories and anticipations). That is, beliefs need not be directly grounded in the "here-and-now" observations.

D. Iterative belief refinement

We now turn our attention to the way stable beliefs can be constructed bottom-up from the initial input provided by the perceptual beliefs. The formation of stable beliefs proceeds in four consecutive steps: (1) *perceptual grouping*, (2) *multi-modal fusion*, (3) *tracking* and (4) *temporal smoothing*. Fig. 3(b) provides a graphical illustration of this process.

1) Perceptual grouping: The first step is to decide which percepts from different modalities belong to the same real-world entity, and should therefore be grouped into a belief. For a pair of two percepts p_1 and p_2, we infer the likelihood of these two percepts being generated from the same underlying entity in the real-world. This is realised by checking whether their respective features *correlate* with each other.

The probability of these correlations are encoded in a Markov Logic Network. The formulae might for instance 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 the features "shape: cylindrical" and "object: ball". Eq. 10 illustrates the correlation between the cylindrical shape (`Cyl`) and the object label "mug" (`Mug`).

$$w_i \quad \exists i,j \;\; \texttt{Shape}(\texttt{x},\texttt{i}) \wedge \texttt{Cyl}(\texttt{i}) \;\wedge \; \texttt{Label}(\texttt{y},\texttt{j}) \wedge \texttt{Mug}(\texttt{j}) \rightarrow \texttt{Corr}_\texttt{i}(\texttt{x},\texttt{y}) \qquad (10)$$

A grouping of two percepts will be given a high probability if one or more feature pairs correlate with each other, and there are no incompatible feature pairs. This process is triggered at each insertion or update of percepts. Its outcome is a probability distribution over possible percept unions.

2) Multi-modal fusion: We want multi-modal beliefs to go beyond the simple superposition of isolated modal contents. Multi-modal information should be *fused*. In other words, the modalities should co-constrain and refine each other, yielding new multi-modal estimations which are globally more accurate than the uni-modal ones. We are not talking here about low-level fusion on a metric space, but about fusion based on conceptual structures. These approaches should be seen as complementary with each other.

Multi-modal fusion is also specified in a Markov Logic Network. As an illustration, assume a multi-modal belief `B` with a predicate $\texttt{Position}(\texttt{B},\texttt{loc})$ expressing the positional coordinates of an entity, and assume the value `loc` can be estimated via distinct modalities a and b by way of two

predicates $\texttt{Position}_{\texttt{(a)}}(\texttt{U},\texttt{loc})$ and $\texttt{Position}_{\texttt{(b)}}(\texttt{U},\texttt{loc})$ included in a percept union U.

$$w_i \quad \texttt{Position}_{\texttt{(a)}}(\texttt{U},\texttt{loc}) \rightarrow \texttt{Position}(\texttt{B},\texttt{loc}) \quad (11)$$

$$w_j \quad \texttt{Position}_{\texttt{(b)}}(\texttt{U},\texttt{loc}) \rightarrow \texttt{Position}(\texttt{B},\texttt{loc}) \quad (12)$$

The weights w_i and w_j specify the relative confidence of the modality-specific measurements.

3) Tracking: Environments are dynamic and evolve over time – and so should beliefs. Analogous to perceptual grouping which seeks to bind observations over modalities, tracking seeks to bind beliefs *over time*. Both past beliefs (memorisation) and future beliefs (anticipation) are considered. The outcome of the tracking step is a distribution over temporal unions, which are combinations of beliefs from different spatio-temporal frames.

The Markov Logic Network for tracking works as follows. First, the newly created belief is compared to the already existing beliefs for similarity. The similarity of a pair of beliefs is based on the correlation of their content (and spatial frame), plus other parameters such as the time distance between beliefs. If two beliefs $\texttt{B}_1$ and $\texttt{B}_2$ turn out to be similar, they can be grouped in a temporal union U whose temporal interval is defined as $[\texttt{start}(\texttt{B}_1), \texttt{end}(\texttt{B}_2)]$.

4) Temporal smoothing: Finally, temporal smoothing is used to refine the estimates of the belief content *over time*. Parameters such as recency have to be taken into account, in order to discard outdated observations.

The Markov Logic Network for temporal smoothing is similar to the one used for multi-modal fusion:

$$w_i \quad \texttt{Position}_{\texttt{(t-1)}}(\texttt{U},\texttt{loc}) \rightarrow \texttt{Position}(\texttt{B},\texttt{loc}) \quad (13)$$

$$w_j \quad \texttt{Position}_{\texttt{(t)}}(\texttt{U},\texttt{loc}) \rightarrow \texttt{Position}(\texttt{B},\texttt{loc}) \quad (14)$$

IV. Extensions

A. Salience modelling

The belief formula of an entity usually contains a specific feature representing its *salience*. The salience value gives an estimate of the "prominence" or quality of standing out of a particular entity relative to neighboring ones. It allows us to drive the attentional behaviour of the agent by specifying which entities are currently in focus.

In our model, the salience is defined as a real-valued measure which combines several perceptual measures such as the object size and its linear and angular distances relative to the robot. During linguistic interaction, these perceptual measures can be completed by measures of linguistic saliency, such as the recency of the last reference to the object.

The salience being real-valued, its probability is defined as a density function $\Re \rightarrow [0, 1]$.

B. Referencing beliefs

Beliefs are high-level symbolic representations available for the whole cognitive architecture. As such, they provide an unified model of the environment which can be used during interaction. An important aspect of this is *reference resolution*, which connects linguistic expressions such as "this box" or "the ball on the floor" to the corresponding beliefs about entities in the environment. Reference resolution is performed via a Markov Logic Network specifying the correlations between the linguistic constraints of the referring expression and the belief features (in particular, the entity saliency and its associated categorical knowledge).

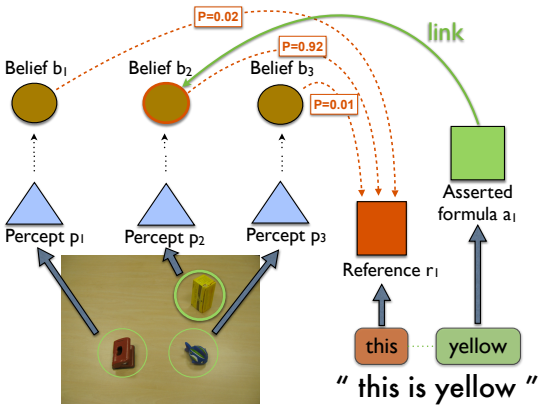


Fig. 4. An utterance such as "This is yellow" illustrates the two mechanisms of referencing and belief extension. First, the expression "this" is resolved to a particular entity. Since "this" is a (proximal) deictic, the resolution is performed on basis of saliency measures. The belief $\texttt{B}_2$ is selected as most likely referent. Second, the utterance also provides new information – namely that the object is yellow. This asserted content must be incorporated into the robot's beliefs. This is done by constructing a new belief which is linked (via a pointer) to the one of the referred-to entity.

Formula (15) illustrates the resolution of a referring expression R containing the linguistic label "mug" to a belief B which includes a label feature with value Mug:

$$w_i \quad \exists i, j \;\; \texttt{Label}(\texttt{B},\texttt{i}) \wedge \texttt{Mug}(\texttt{j}) \wedge \texttt{Ref}(\texttt{R},\texttt{j}) \wedge \texttt{Mug}(\texttt{j}) \rightarrow \texttt{Resolve}(\texttt{R},\texttt{B}) \quad (15)$$

The resolution process yields a probability distribution over alternative referents, which is then retrieved by the communication subsystem for further interpretation.

C. Asserting new information

In Section III-D, we described how beliefs can be formed from percepts, bottom-up. When dealing with cognitive robots able to reflect on their own experience, anticipate possible events, and communicate with humans to improve their understanding, beliefs can also be manipulated "top-down" via high-level cognitive functions such as reasoning, planning, learning and interacting.

We concentrate here on the question of belief extension via interaction. In addition to simple reference, interacting with a human user can also provide *new* content to the beliefs. Using communication, the human user can directly extend the robot's current beliefs, in a top-down manner, without altering the incoming percepts. The epistemic status of this information is *attributed*. If this new information conflicts with existing knowledge, the agent can decide to trigger a clarification request to resolve the conflict.

Fig. 4 provides an example of reference resolution coupled with a belief extension.

D. Belief filtering

Techniques for *belief filtering* are essential to keep the system tractable. Given the probabilistic nature of the framework, the number of beliefs is likely to grow exponentially over time. Most of these beliefs will have a near-zero probability. A filtering system can effectively prune such unecessary beliefs, either by applying a minimal probability threshold on them, or by maintaining a fixed maximal number of beliefs in the system at a given time. Naturally, a combination of both mechanisms is also possible.

In addition to filtering techniques, *forgetting* techniques could also improve the system efficiency [4].

V. Conclusion

In this paper, we presented a new approach to the construction of *rich belief models* for situation awareness. These beliefs models are spatio-temporally framed and include epistemic information for multi-agent settings. Markov Logic is used as a unified representation formalism, allowing us to capture both the complexity (relational structure) and uncertainty (partial observability) of typical HRI domains.

The implementation of the approach outlined in this paper is ongoing. We are using the Alchemy software[3] for efficient probabilistic inference. The binder system revolves around a central working memory where percepts can be inserted, modified or deleted. The belief model is automatically updated to reflect the incoming information.

Besides the implementation, future work will focus on three aspects. The first aspect pertains to the use of *machine learning techniques* to learn the model parameters. Using statistical relational learning techniques and a set of training examples, it is possible to learn the weights of a given Markov Logic Network [17]. The second aspect concerns the extension of our approach to non-indexical epistemic knowledge –i.e. the representation of *events*, *intentions*, *plans*, and *general knowledge* facts. Finally, we want to evaluate the empirical performance and scalability of our approach under a set of controlled experiments.

VI. Acknowledgments

The research leading to these results has received funding from the European Union's 7th Framework Programme [FP7/2007-2013] under grant agreement No. 215181 (CogX). The authors wish to thank Miroslav Janíček and Hendrik Zender for useful discussions.

References

[1] L. W. Barsalou. Abstraction in perceptual symbol systems. *Philosophical Transactions of the Royal Society of London: Biological Sciences*, 358:1177–1187, 2003.

[2] H. H. Clark and E. F. Schaefer. Contributing to discourse. *Cognitive Science*, 13:259–294, 1989.

[3] F. Doshi and N. Roy. Spoken language interaction with model uncertainty: an adaptive human-robot interaction system. *Connection Science*, 20(4):299–318, 2008.

[4] S. T. Freedman and J. A. Adams. Human-inspired robotic forgetting: Filtering to improve estimation accuracy. In *Proceedings of the 14th IASTED International Conference on Robotics and Applications*, pages 434–441, 2009.

[5] D. L. Hall and J. Llinas. An introduction to multisensor data fusion. *Proceedings of the IEEE*, 85(1):6–23, August 2002.

[6] S. Harnad. The symbol grounding problem, 1990.

[7] N. Hawes and J. Wyatt. Engineering intelligent information-processing systems with cast. *Advanced Engineering Infomatics*, To Appear.

[8] N. Hawes, H. Zender, K. Sjöö, M. Brenner, G.-J. M. Kruijff, and P. Jensfelt. Planning and acting with an integrated sense of space. In *Proceedings of the 1st International Workshop on Hybrid Control of Autonomous Systems – Integrating Learning, Deliberation and Reactive Control (HYCAS)*, pages 25–32, Pasadena, CA, USA, July 2009.

[9] H. Jacobsson, N.A. Hawes, G.-J. M. Kruijff, and J. Wyatt. Crossmodal content binding in information-processing architectures. In *Proceedings of the 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Amsterdam, The Netherlands, March 12–15 2008.

[10] D. Koller, N. Friedman, L. Getoor, and B. Taskar. Graphical models in a nutshell. In L. Getoor and B. Taskar, editors, *Introduction to Statistical Relational Learning*. MIT Press, 2007.

[11] G-J. M. Kruijff, J.D. Kelleher, and N. Hawes. Information fusion for visual reference resolution in dynamic situated dialogue. In *Perception and Interactive Technologies (PIT 2006)*. Spring Verlag, 2006.

[12] G.-J. M. Kruijff, John D. Kelleher, and N. Hawes. Information fusion for visual reference resolution in dynamic situated dialogue. In *Perception and Interactive Technologies: International Tutorial and Research Workshop, PIT 2006*, volume 4021 of *Lecture Notes in Computer Science*, pages 117 – 128, Kloster Irsee, Germany, June 2006. Springer Berlin / Heidelberg.

[13] I. Kruijff-Korbayová, S. Ericsson, K. J. Rodríguez, and E. Karagjosova. Producing contextually appropriate intonation in an information-state based dialogue system. In *EACL '03: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*, pages 227–234, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[14] P. Lison. Robust processing of situated spoken dialogue. In *Von der Form zur Bedeutung: Texte automatisch verarbeiten / From Form to Meaning: Processing Texts Automatically*. Narr Verlag, 2009. Proceedings of the GSCL Conference 2009 , Potsdam, Germany.

[15] P. Lison and G.-J. M. Kruijff. Salience-driven contextual priming of speech recognition for human-robot interaction. In *Proceedings of ECAI 2008*, Athens, Greece, 2008.

[16] H. Poon and P. Domingos. Sound and efficient inference with probabilistic and deterministic dependencies. In *AAAI'06: Proceedings of the 21st national conference on Artificial intelligence*, pages 458–463. AAAI Press, 2006.

[17] M. Richardson and P. Domingos. Markov logic networks. *Machine Learning*, 62(1-2):107–136, 2006.

[18] D. Roy. Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence*, 167(1-2):170–205, 2005.

[19] D. Roy and E. Reiter. Connecting language to the world. *Artificial Intelligence*, 167(1-2):1–12, 2005.

[20] M. Stone and R.H. Thomason. Context in abductive interpretation. In *Proceedings of EDILOG 2002: 6th workshop on the semantics and pragmatics of dialogue*, 2002.

[21] A. Vrečko, D. Skočaj, N. Hawes, and A. Leonardis. A computer vision integration model for a multi-modal cognitive system. In *IEEE/RSJ International Conference on Intelligent RObots and Systems*, pages 3140–3147, 2009.

[22] M. Walker, S. Whittaker, A. Stent, P. Maloor, J. Moore, M. Johnston, and G. Vasireddy. Generation and evaluation of user tailored responses in multimodal dialogue. *Cognitive Science*, 28(5):811–840, 2004.

[23] J. Williams and S. Young. Partially observable markov decision processes for spoken dialog systems. *Computer Speech and Language*, 21(2):231–422, 2007.

[24] H. Zender, G.-J. M. Kruijff, and I. Kruijff-Korbayová. Situated resolution and generation of spatial referring expressions for robotic assistants. In *Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence (IJCAI-09)*, pages 1604–1609, Pasadena, CA, USA, July 2009.

[3] Cf. `http://alchemy.cs.washington.edu/`