Combining World and Interaction Models for Human-Robot Collaborations

Cynthia Matuszek* Andrzej Pronobis* Luke Zettlemoyer Dieter Fox
(cynthia | pronobis | lsz | fox @ cs.washington.edu)

Abstract

As robotic technologies mature, we can imagine an increasing number of applications in which robots could soon prove to be useful in unstructured human environments. Many of those applications require a natural interface between the robot and untrained human users or are possible only in a human-robot collaborative scenario. In this paper, we study an example of such scenario in which a visually impaired person and a robotic "guide" collaborate in an unfamiliar environment. We then analyze how the scenario can be realized through language- and gesture-based human-robot interaction, combined with semantic spatial understanding and reasoning, and propose an integration of semantic world model with language and gesture models for several collaboration modes. We believe that this way practical robotic applications can be achieved in human environments with the use of currently available technology.

Introduction

As robots become more capable of performing complex tasks, there are an increasing number of scenarios and environments in which they may be deployed usefully. However, before we can build practical, useful domestic or personal robots, a number of challenges and shortcomings in the current technology must be addressed. Current designs often trade the complexity of the systems and scenarios for overall robustness. As a result, robots can explore only portions of typical environments and often do not attempt to manipulate objects at the level required for many realistic applications. As well, interfaces to current systems are often very limited, requiring the user to learn how to interact with a robot, rather than the system learning how the user naturally communicates.

In this paper, we place the focus on human-robot interaction. We claim that enabling robots to communicate and collaborate with humans in a natural way is critical to making them practical for realistic applications. The reasoning is twofold: first, many of the envisioned applications require a natural interface between the robot and untrained human users; and second, many of the existing problems with robotic technologies can be overcome by allowing the robots to collaborate with humans and share competences.

Following this principle, we propose a realistic scenario in which human-robot interaction and collaboration could be used to enable a practical robotic application that uses currently available technologies. Our scenario describes a sequence of interactions between a visually impaired person and a robotic "guide" in an unfamiliar environment. In this case, both the robot and the person have capabilities that are unavailable to the other agent, and the collaboration results in a clear mutual benefit.

This example application calls out a number of necessary technical components, many of which are spread across distinct research areas; in this work we focus on two: language- and gesture-based multimodal human-robot interaction, and spatial understanding and reasoning, both of which have proven to be independently useful in realistic demonstrations. At a high level, the approach we take to integrating these components is based on interaction of two problem-specific multimodal representations, namely, a semantic world model and a human-robot interaction model. We analyze the scenario from the perspective of those two components and their integration, and identify several collaboration modes, which we study in more detail.

Related Work

Semantic understanding of the world is crucial for generalizing about environments and objects, interacting with human users, and so on. As a result, the *semantic mapping problem* has, in recent years, received significant attention (Galindo et al. 2005; Zender et al. 2008; Vasudevan and Siegwart 2008; Meger et al. 2008; Tenorth et al. 2010). None of these methods uses topology of the environment or general appearance of places as a source of semantic information, although there is large body of work on appearance-based place categorization. Most systems primarily use objects for extracting spatial semantics; we will build on (Pronobis and Jensfelt 2012), which uses all of general appearance and geometry of places, object information, topological structure, and human input to create semantic maps.

We will also use the *conceptual map* from (Pronobis and Jensfelt 2012), in which a probabilistic ontology of the indoor environment is discovered and linked to instances of spatial entities for inference. This is most similar to this of (Galindo et al. 2005) and (Zender et al. 2008); however, those ontologies are built manually and use traditional AI reasoning techniques which do not incorporate the uncertainty that is inherently connected with semantic information obtained through robot sensors in realistic environments.

Robots deployed in unconstrained real-world settings will need to learn to understand the intentions and references of users, from the users themselves. In order to interact naturally and smoothly with end users in an unconstrained way, it is necessary to understand human inputs (here, natural language commands and queries, and indicative gestures). While gesture recognition has been extensively explored for use in such interfaces (Mitra and Acharya 2007), our focus

*The first two authors contributed equally to this paper.

is on using non-scripted gestures to control physical systems. (Matuszek et al.)

The language component of our work falls into the class of *grounded language acquisition* approaches, in which language is learned from situated context. Learning about the world and human communication from a combination of language and sensor data has achieved particular success in understanding commands, for example in navigation (Hawes and Wyatt 2010), robot command interpretation (Tenorth and Beetz 2012), localization (Spexard et al. 2006), and search and rescue (Cantrell et al. 2012).

There has been significant work on using learning to induce probabilistic semantic parsers. Parsed natural language can be grounded in a robots world and action models, taking perceptual and grounding uncertainty into account, enabling instruction-following in such tasks as robot direction-following and operation (Matuszek et al. 2012b; Tellex et al. 2012). In this paper we propose an extension to our work on using semi-supervised learning to jointly learn models of language and vision, (Matuszek* et al. 2012a) which builds on work on supervised learning of Combinatory Categorial Grammar (CCG) lexicons for semantic parsing. (Kwiatkowski et al. 2011; Artzi and Zettlemoyer 2013)

Integrated robotic systems consisting of a large number of independent components still pose a great challenge in practical applications, but even in this arena there exist examples of recent successful work (Guizzo and Ackerman 2012). Work on natural language grounding (Matuszek et al. 2012b), unconstrained gesture comprehension (Matuszek et al.), spatial understanding and semantic mapping (Pronobis 2011), assistive robotics (Kulyukin 2006), and natural user interfaces (Wigdor and Wixon 2011) – when integrated into a system that takes advantage of their capabilities – offer hope of a robotic system that can be robustly and usefully deployed. Some of the remaining gap can be bridged by thinking of robots as *collaborators*, with capabilities that supplement and are supplemented by human abilities. (Veloso et al. 2012)

Human-Robot Collaborative Scenario

Many believe that we will see large growth in the areas of service and assistive robotics in upcoming years, areas where even limited robotic intervention has the potential to offer substantial benefits. In this section, we describe a possible assistive technology scenario that draws in work in human/robot interaction, vision, semantic mapping, and natural language grounding, among others.

The scenario is deliberately designed to both take advantage of work to date, and to show the potential of near-term technology in supporting natural, useful robotic systems. We annotate the scenario according to different types of interaction; after presenting the scenario, we discuss the components necessary to support such a system. In this collaboration, the robot uses interactions and its world model to roughly model a human's belief state, rather than maintaining such a model explicitly as in (Breazeal et al. 2005). While limited, this model of the users' cognitive state is sufficient to allow for performance of certain useful tasks.

We divide the interactions into types of high-level interaction, which we label in the dialog as follows. 'Goal' refers to an explicit statement by a person of a goal to be carried out; 'Clarify' is an action in which the robot seeks clarification of the world-state from a person; 'Respond' means a robot's response to a human query about the world; 'Explain' is an interaction in which a human clarifies or adjusts the robot's belief state; 'Inform' is a standing goal for the robot to inform a user about anything atypical or interesting, or informs the person about an action being taken; and a 'Request' occurs when the robot needs assistance from a human. The section on collaboration modes explores the first four in more detail.

Scenario: Human-Robot Collaboration

Robot and person enter office building

Default goals: Inform person of anything atypical or interesting + Gather knowledge

Robot: I'm in a new building. I think we are in an office building, in a big hall. [Inform]
Person: Let's find the coffee room on the second floor. [Goal]
Robot: What is a coffee room? [Explain]
Person: It's like a kitchen with a coffee machine. [Explain]
Robot: Okay. [Explain]
Goal: Navigate to coffee room on second floor
Robot: Which floor are we on? [Clarify]
Person: The ground floor. [Clarify]
Robot: Ok. [Clarify]
Robot: Looking for an elevator. [Inform]
Robot: There seem to be a lot of tables in the hall. [Inform]
Robot: We are at the elevators. [Inform]
Robot: Could you summon the elevator? [Request]
Person: Where is the elevator button? [Respond]
Additional goal: take user to place containing the elevator button
Robot: I will take you. *(leads)* [Respond]

Events unfold. Robot and human are now in the coffee room.

Goal: default
Person: Is this decaf? *(points)* [Respond]
Robot: What does a decaf look like? [Explain]
Person: It would have a green lid. [Explain]
Robot: Yes, it is a decaf. [Respond]

This use case, while complex, is not unrealistically far beyond the current state of the art. However, realizing it will require integrating components from a number of areas: natural language understanding, vision, semantic environment mapping, knowledge representation, planning and cost analysis, and natural language generation, among others.

In the remainder of this paper, we discuss the technology underpinning the semantic mapping and human interaction components in more detail. We assume the remaining roles will be filled by work performed in those fields, although in practice an initial implementation would likely rely on simpler approaches (such as templated language generation).

Semantic World Model

Our world model is a holistic representation of complex, cross-modal, uncertain spatial knowledge. It includes knowledge about spatial topology, the presence of objects (e.g. cups or bottles) and landmarks (e.g. a door or a shelf),

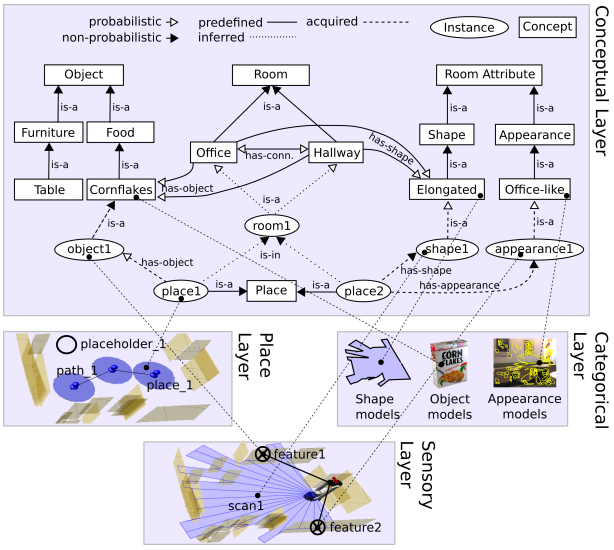


Figure 1: The layered structure of the spatial representation and an excerpt of the ontology of the conceptual layer. The conceptual layer comprises knowledge about concepts (rectangles), relations between concepts, and instances of spatial entities (ellipses).

object attributes (e.g. shape or color) as well as room attributes such as shape (e.g. elongated or rectangular), size or general appearance (e.g. kitchen-like or corridor-like). This representation follows the principles presented in (Pronobis et al. 2010; Pronobis and Jensfelt 2012) and can be built in real-time using their semantic mapping algorithm. It abstracts multi-modal sensory information and integrates it with conceptual common-sense knowledge in a fully probabilistic fashion. This keeps the representations compact, makes knowledge robust to dynamic changes, and permits reasoning about concepts that are not directly observed.

The structure of the representation is presented in Fig. 1. The framework comprises four layers, each focusing on a different level of knowledge abstraction, from low-level sensory input to high-level conceptual symbols. The lowest level of the representation is the sensory layer, which maintains an accurate representation of the robot's environment. Above, the place layer discretizes the continuous space into *places*. Places connect to other places by *paths,* which are generated as the robot travels between them, forming a topological map. The categorical layer comprises universal categorical models, which describe objects and landmarks, as well as room and object attributes such as geometrical models of room shape or visual models of appearance.

The highest-level layer is the conceptual layer, populated by instances of spatial concepts and providing a unified representation relating sensed instance knowledge from lower-level layers to general common-sense conceptual knowledge. Moreover, it includes a taxonomy of human-compatible spatial concepts. It is the conceptual layer which would contain the information that kitchens commonly contain refrigerators and have certain appearance; this allows the robot to make inferences about the environment (e.g., the presence of a cereal box makes it more likely that the current room is a kitchen.

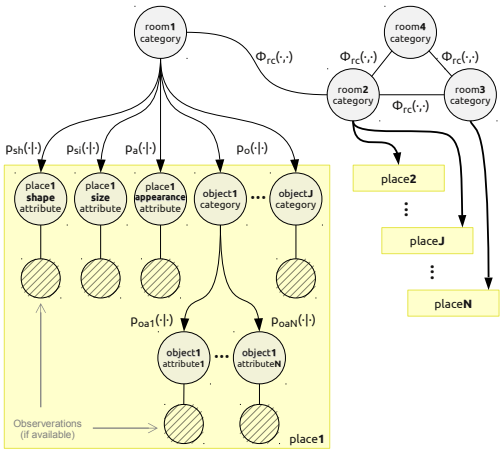


Figure 2: Structure of the chain graph model of the conceptual map. Vertices represent random variables; edges show probabilistic relationships among random variables. Textured vertices show observations corresponding to sensed evidence.

The concept of *a room* is exploited in the conceptual layer in order to group locations. Rooms tend to share similar functionality and semantics and are typically assigned semantic categorical labels e.g. a double office. This make them appropriate units for knowledge integration.

A visualization of the data representation of the conceptual layer is shown in Fig. 1. This representation is *relational*, describing common-sense knowledge as relations between concepts (e.g. `kitchen has-object cornflakes`), and describes instance knowledge as relations between either instances and concepts (e.g. `object1 is-a cornflakes`), or instances and other instances (e.g. `place1 has-object object1`).

Relations in the conceptual layer are *predefined*, *acquired* from observations, or *inferred* from the conceptual layer, and can either be deterministic or probabilistic. Probabilistic relations allow the expression of statistical dependencies and uncertainty as in the case of the `kitchen has-object cornflakes` or `room1 is-a hallway` relations which hold only with a certain probability.

The Conceptual Map

In order to allow probabilistic inference in the conceptual layer, it is compiled into a *chain graph* representation (Lauritzen and Richardson 2002), which we refer to as a *conceptual map* (Pronobis and Jensfelt 2012). Chain graphs allow us to model both "directed" causal relations (such as `is-a`) as well as "undirected" symmetric or associative relations (such as room connectivity). The structure of the conceptual map is adapted on the fly, reflecting the state of the underlying topological map and observations gathered by the robot.

The structure of the conceptual map is presented in Fig. 2. Each discrete place instance is represented by a set of random variables describing object and landmark instances and spatial attributes associated with that place. These variables are connected to a random variable describing the functional category of the room. The distributions over the values of those variables represent the `is-a` relation in Figure 1. The

distributions $p_{sh}(\cdot|\cdot)$, $p_{si}(\cdot|\cdot)$, $p_a(\cdot|\cdot)$, $p_o(\cdot|\cdot)$, $p_l(\cdot|\cdot)$ represent common sense knowledge about the relations between room categories and room shape, size, and appearance attributes, as well as categories of objects and landmarks that are typically present. Additional variables represent information about object attributes such as color or shape and the distributions $p_{oa1}(\cdot|\cdot)$, ..., $p_{oaN}(\cdot|\cdot)$ encode the common sense knowledge relating objects of certain categories with their attributes. Moreover, the room category variables are connected by undirected links to one another according to the topological map. The potential functions $\phi_{rc}(\cdot,\cdot)$ describe knowledge about typical connectivity of rooms of certain categories (e.g., kitchens are more likely to be connected to corridors than to other kitchens).

The variables describing room attributes, objects, landmarks and object attributes can be linked to observations gathered by the robot in the form of features extracted directly from the sensory input. As proposed in (Pronobis and Jensfelt 2012), these links (textured vertices in Fig. 2) can be quantified by categorical models of sensory information. This common-sense knowledge about room connectivity, shapes, sizes and appearances can be acquired by analyzing annotations of existing databases, typically used for experiments with place categorization (Pronobis and Jensfelt 2012; Pronobis and Caputo 2009). As shown in (Hanheide et al. 2011), the correspondences between object and landmark categories and certain functional categories of rooms can be obtained by exploiting common-sense knowledge databases or by analyzing results of image search engines.

Reasoning about Unexplored Space

Having a probabilistic relational conceptual representation allows us to perform uncertain inference about concepts based on their relations to other concepts, as well as based on direct observations; this permits spatial reasoning about unexplored space. Consider the case of predicting the presence of objects of certain categories in a room with a known category. This can be easily performed in our model by adding variables and relations for object categories without providing the actual object observations. As shown in (Aydemir et al. 2013), this can be exploited to continuously predict the existence of objects based on other semantic cues.

Another way of using the predictive power of the conceptual map is to predict the existence of a room of a certain category in the unexplored space. In this case, the conceptual map is extended from the room in which unexplored space exists with variables representing categories of hypothesized rooms for different possible room configurations in the unexplored space. For each configuration, the categories of the hypothesized rooms are calculated and the obtained probabilities of existence of rooms of certain categories as well as room attributes or objects potentially present in the hypothesized rooms are summed over all possible configurations. For details on real-time implementations, see (Pronobis and Jensfelt 2012; Aydemir et al. 2013).

Language & Gesture

Our motivating scenario relies heavily on successfully understanding a user's unscripted, natural input. This is intentional; as robots move into the real world, the importance of enabling untrained users to interact with them in a natural way can only increase. Unfortunately, human interactions can be quite complex, even in a constrained task like "indicating objects" (as in in Fig. 5). Here, someone referring to a set of blocks as the "blue red *(sic)* rectangles", while sketching a rough circle above them, is comprehensible to a person, but outside the scope of current user interfaces – a situation which will become harder to accept as robots become more widely deployed and capable.

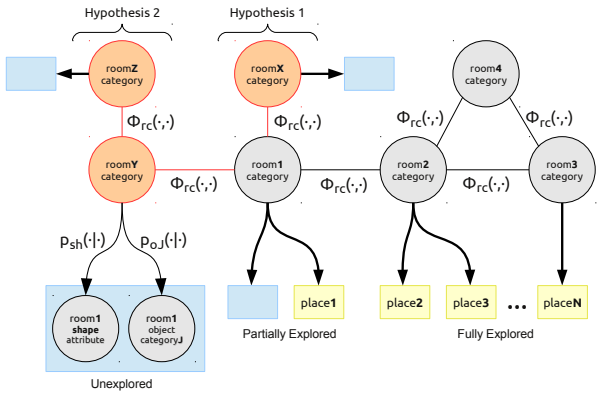


Figure 3: Examples of extensions of the conceptual map permitting reasoning about unexplored space.

In this section, we describe work on understanding two human input modalities that appear in our example scenario. We first briefly describe work on understanding *indicative gestures*, such as pointing, which are intended to call attention to something tangible. We then describe a language model suitable for interpreting the utterances.

Gesture Understanding

Because it is often natural to use gesture to direct attention (Topp et al. 2006), understanding such *indicative gestures* is an appropriate first step for the proposed system. Compare, for example, "Put the mug that's on the table in the cupboard to the left of the stove," versus "Put this mug in there." The latter is a natural way of communicating a need to a robot, set in a context where traditional input devices such as keyboards are lacking.

Existing multimodal interfaces generally require the user to learn how a system expects to be instructed, rather than the system learning how the user naturally communicates. Current gestural interfaces have primarily focused on gesture *recognition* – that is, on identifying a lexicon of gestures which the user must first learn. (Malizia and Bellucci 2012) Instead, we propose to learn to recognize (possibly user-specific) gestures in the same fashion as (Matuszek et al.). The *acquired* knowledge that an object or landmark is the subject of `object-indicated` event is then integrated into the world model as shown in Fig. 6.

Here, a depth camera is used to extract spatial features, then sparse coding (Yang et al. 2009) is used to learn Hierarchical Matching Pursuit features (Bo, Ren, and Fox 2011) suitable for the binary `object-indicated` classification

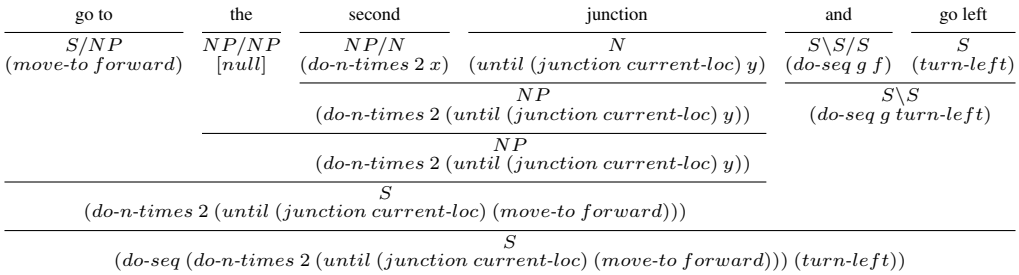
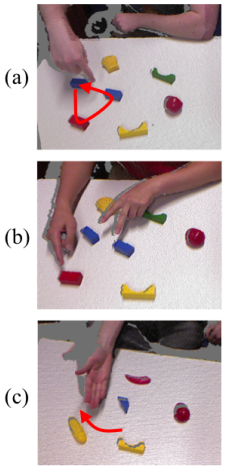


Figure 4: CCG parse of a test sentence performed by the learned parser. Here the natural language input is first, followed by alternating syntactic categorization and λ-calculus logical forms. The bottom row shows the final representation that will be incorporated into the world model as a user-provided goal.



(d)

"Three rectangles. These three. Two of them are blue. Two blue red rectangles and one red rectangle."

"Okay now I have three objects. One two three. They both they all three have the same… shape. Like little thick sticks with some rounded corners. But those two are blue and this one is red."

"Please – we need the three elements here... that... that all look the same."

(e)

"The yellow ones."

"Okay, so these are two yellow objects. One of that – is – corn-like. Longs, long elliptic object. and the other one is like rectangle with the ... with an arc cut out."

Figure 5: Examples of unscripted references to objects. (a) A circular pointing motion; (b) pointing with multiple fingers and both hands; (c) an open-handed sweep above objects. (d) and (e) give examples of different language for the two scenarios shown.

task. Specifically, a sequence of vectors of spatial features is extracted and coded as sparse linear combinations of codewords selected from a codebook (Lai et al. 2013), which is learned from example gestures. Data can then be represented by a sparse, linear combination of these codebook entries. Logistic regression over this sequence of vectors yields an indicated/not indicated classification, which can then be integrated into the system's beliefs.

Language Model

We treat understanding the language present in the human interaction as a class of *language grounding* – using a physically grounded setting to provide information for connect language and percepts. In order to understand language about physical systems, we extract semantically meaningful representations of language content by *parsing*, then map those representations to the world model.

For this work, parsing is performed using an extended version of the Unification-Based Learner, UBL (Kwiatkowski et al. 2010). The grammatical formalism used by UBL is a probabilistic version of *combinatory categorial grammars*, or CCGs (Steedman 2000), a type of phrase structure grammar. CCGs model both the syntax (language constructs such as NP for noun phrase) and the semantics (expressions in λ-calculus) of a sentence. UBL creates a parser by inducing a probabilistic CCG (PCCG) from a set of training examples. PCCG-based algorithms are able to efficiently generate n-best parses, allowing for jointly considering a parse model and a world model derived from sensor data; multiple parse probabilities can be combined with the world model, allowing the combined system to choose the language interpretation that makes the most sense given the context.

This approach has proven effective in interpreting input for understanding imperatives (Matuszek et al. 2012b); Fig. 4 shows how a navigation imperative can be parsed into the formal representation underpinning our world model, in a manner similar to understanding statements such as "find the kitchen." The same system has been used for understanding descriptions of previously unfamiliar world objects (Matuszek* et al. 2012a), and a similar but simpler language model has been combined with world knowledge to interpret indication speech ((Matuszek et al.)) – that is, to determine whether a person is describing a particular object. Intuitively, an object o has been "referred to" if it has attributes the user mentions; if parsed language contains a reference to the attribute `green-color`, and acquired world knowledge includes `(has-color object1 green-color)`, then `object1` is a likely referral target.

Collaboration Modes

Understanding Explicit Goal Statements Interactions marked as 'Goal's in our scenario are focused on interpreting human statements that explicitly state location and navigation goals for the robot. To interpret such an utterance, the language parsing component first generates several parses representing semantic interpretations, each associated with a measure of uncertainty. From these, the system extracts relevant spatial entities, the existance of which should be evaluated in the world model. For example, given the sentence "Let's find the coffee room on the second floor," one possible interpretation is: `(do-until (and (is-room-type current-loc coffee-room) (has-attribute current-loc level) (equals level 2)) (create-event (explore-forward)))`.

Given this parse, the system is aware that we care about a coffee room, and that it is on the second floor. Using the world model constructed for a partially explored en-

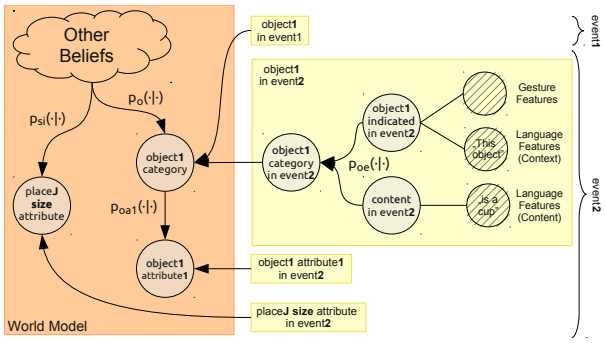


Figure 6: Integration of language/gesture models and world model.

vironment, we produce the probability that there exists a world structure that matches the goal. This can be done by combining the knowledge about already explored space and the predictions for yet unexplored space as described in Section . Additional clarifying dialog might yield the belief `(subtype coffee-room kitchen)`. Inference is performed by combining the probabilities inferred for room category and room property variables to achieve the probability of an event such that, over all places in unexplored space, there exists a place of category room and type kitchen.

Clarifying Robot Belief State The purpose of the 'Clarify' collaboration mode is to allow the robot to increase the certainty in its internal beliefs about the world by using a human as a source of information. Such interaction might require careful planning and trading the benefits and costs of a dialogue with a human versus using direct robot perception. This topic has been explored before in the context of planning (Hanheide et al. 2011). Here, we analyze the problem from the point of view of the world model and its interaction with language and gesture understanding. The interaction consists of two elements:

1. *Uncertainty-driven query* initiated by the robot, e.g. "What room are we in?" or "What floor are we on?"
2. *Assertion about the world state* provided by the human, e.g. "We are in a kitchen" or "on the ground floor."

The first step is to realize that there is a gap in the robot's knowledge, in order to initiate the dialogue. Such gaps in knowledge can be identified based on the probability of the most likely concept associated with the spatial entity in the world model. This in turn, might result in the robot generating a templated language query based on the taxonomy of concepts associated with the spatial entity, e.g. "What is the size of this room?" for the room size attribute.

In order to incorporate the human assertion into the world state, we propose the integration between the world model and the language and gesture models shown in Figure 6. For every human-robot interaction event and a potential spatial entity of interest, we generate a set of random variables (enclosed in a yellow rectangle). First, we use the gesture and language understanding to infer whether the spatial entity was indicated during the event (e.g. the user pointed at an object and said "This"). We combine this information with the inferred content of the assertion extracted using the language model (e.g. "is a cup"). This allows us to reason about the influence of the event on the value of the variable describing the spatial entity. In order to represent the uncertainty in that association, similar variables can be created for the same event and other spatial entities which were potentially indicated.

Responding to Human World-state Query Knowing what a user should be informed of is a difficult problem. Having a robot continuously explain everything about the world state is obviously impractical; however, the human may request specific information. In this mode, the robot should 'Respond' effectively but concisely.

The first step is to parse a question correctly: "Is this decaf?" calls for a different type of response than "Where is the elevator button?" Once the robot has committed to a parse, possibly after asking clarifying questions, the answer and the appropriate mechanism for providing it can be chosen by analyzing object category variables. While the first question can be satisfied by a spoken response, the internal representation of a location may be difficult to convey in a linguistically appropriate frame of reference, causing the robot to respond by adding a `lead-to-place` action.

Clarifying Human Input In the 'Explain' interaction, the robot seeks assistance from a user on understanding input. We discuss the case of input containing an unfamiliar word (here, 'decaf' may be a previously non-encountered abbreviation). If the parser partially succeeds, the robot can be aware that a sentence requires a noun phrase for completion. This provides enough information to induce a templated response, e.g., "What is a decaf?" –which, while not perfectly grammatical, is comprehensible.

When the user responds by identifying attributes the object may have (e.g., "It would have a green lid,") , the parsing and disambiguation of (Matuszek* et al. 2012a) can be invoked, in which hypothetical interpretations of the world are combined with beliefs about the world state to select objects. In the context of this integration, this entails, for the object o, redefining the probability distribution describing the dependency between object of a certain category and its attributes (p_{oaI}, $I =$ `green-color`).

Conclusions

In this work, we addressed the problem of enabling realistic robotic applications through human-robot collaboration. We built on top of two existing components, both of which were implemented on robotic hardware employing standard sensors and demonstrated in realistic scenarios. We analyzed a scenario in which human-robot collaboration is crucial and showed how various collaboration modes present in this scenario can be realized using the intersection of a semantic world model and language/gesture interaction models. Our analysis indicate that such scenario is within the reach of current robotic technologies; we intend to continue the work on integrating perception, reasoning and interaction in order to demonstrate the proposed solutions in practice.

Acknowledgements

The authors were supported by grants from the Intel Science and Technology Center for Pervasive Computing and Toyota Infotechnology Inc.

References

Artzi, Y., and Zettlemoyer, L. 2013. Weakly supervised learning of semantic parsers for mapping instructions to actions. *Transactions of the Association for Computational Linguistics* 1(1):49–62.

Aydemir, A.; Pronobis, A.; Göbelbecker, M.; and Jensfelt, P. 2013. Active visual search in unknown environments using uncertain semantics. *IEEE Transactions on Robotics*. (To appear).

Bo, L.; Ren, X.; and Fox, D. 2011. Hierarchical Matching Pursuit for image classification: architecture and fast algorithms. In *Advances in Neural Information Processing Systems*.

Breazeal, C.; Kidd, C. D.; Thomaz, A. L.; Hoffman, G.; and Berlin, M. 2005. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *Intelligent Robots and Systems*. IEEE.

Cantrell, R.; Talamadupula, K.; Schermerhorn, P.; Benton, J.; Kambhampati, S.; and Scheutz, M. 2012. Tell me when and why to do it!: Run-time planner model updates via natural language instruction. In *Proc. of the Conference on Human-Robot Interaction*.

Galindo, C.; Saffiotti, A.; Coradeschi, S.; Buschka, P.; Fernández-Madrigal, J. A.; and González, J. 2005. Multi-hierarchical semantic maps for mobile robotics. In *Proc. of the 2005 International Conference on Intelligent Robots and Systems*.

Guizzo, E., and Ackerman, E. 2012. The rise of the robot worker. *Spectrum, IEEE* 49(10):34–41.

Hanheide, M.; Gretton, C.; Dearden, R. W.; Hawes, N. A.; Wyatt, J. L.; Pronobis, A.; Aydemir, A.; Göbelbecker, M.; and Zender, H. 2011. Exploiting probabilistic knowledge under uncertain sensing for efficient robot behaviour. In *Proc. of the International Joint Conference on Artificial Intelligence*.

Hawes, N. A., and Wyatt, J. L. 2010. Engineering intelligent information-processing systems with CAST. *Advanced Engineering Informatics* 24(1):27–39.

Kulyukin, V. A. 2006. On natural language dialogue with assistive robots. In *Proc. of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, 164–171. ACM.

Kwiatkowski, T.; Zettlemoyer, L.; Goldwater, S.; and Steedman, M. 2010. Inducing probabilistic CCG grammars from logical form with higher-order unification. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing*.

Kwiatkowski, T.; Zettlemoyer, L.; Goldwater, S.; and Steedman, M. 2011. Lexical generalization in CCG grammar induction for semantic parsing. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing*.

Lai, K.; Bo, L.; Ren, X.; and Fox, D. 2013. RGB-D object recognition: Features, algorithms, and a large scale benchmark. In Fossati, A.; Gall, J.; Grabner, H.; Ren, X.; and Konolige, K., eds., *Consumer Depth Cameras for Computer Vision: Research Topics and Applications*. Springer. 167–192.

Lauritzen, S., and Richardson, T. 2002. Chain graph models and their causal interpretations. *J. of Royal Statistical Society* 64(3).

Malizia, A., and Bellucci, A. 2012. The artificiality of natural user interfaces. *Communications of the ACM* 55(3):36–38.

Matuszek, C.; Bo, L.; Zettlemoyer, L.; and Fox, D. Toward Unconstrained Gesture and Speech User Interfaces. (In submission).

Matuszek*, C.; FitzGerald*, N.; Zettlemoyer, L.; Bo, L.; and Fox, D. 2012a. A joint model of language and perception for grounded attribute learning. In *Proc. of the 2012 Int'l Conference on Machine Learning*.

Matuszek, C.; Herbst, E.; Zettlemoyer, L.; and Fox, D. 2012b. Learning to parse natural language commands to a robot control system. In *Proc. of the 13th Int'l Symposium on Experimental Robotics (ISER)*.

Meger, D.; Forssen, P.-E.; Lai, K.; Helmer, S.; McCann, S.; Southey, T.; Baumann, M.; Little, J. J.; and Lowe, D. G. 2008. Curious George: An attentive semantic robot. *Robotics and Autonomous Systems* 56(6).

Mitra, S., and Acharya, T. 2007. Gesture recognition: A survey. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* 37(3):311–324.

Pronobis, A., and Caputo, B. 2009. COLD: The CoSy localization database. *The International Journal of Robotics Research (IJRR)* 28(5):588–594.

Pronobis, A., and Jensfelt, P. 2012. Large-scale semantic mapping and reasoning with heterogeneous modalities. In *Proc. of the 2012 International Conference on Robotics and Automation*.

Pronobis, A.; Sjöö, K.; Aydemir, A.; Bishop, A. N.; and Jensfelt, P. 2010. Representing spatial knowledge in mobile cognitive systems. In *Proc. of Int'l Conference Intelligent Autonomous Systems*.

Pronobis, A. 2011. *Semantic Mapping with Mobile Robots*. Ph.D. Dissertation, KTH Royal Institute of Technology, Stockholm, Sweden.

Spexard, T.; Li, S.; Wrede, B.; Fritsch, J.; Sagerer, G.; Booij, O.; Zivkovic, Z.; Terwijn, B.; and Kröse, B. 2006. BIRON, where are you? Enabling a robot to learn new places in a real home environment by integrating spoken dialog and visual localization. In *Proc. of the 2006 Int'l Conference on Intelligent Robots and Systems*.

Steedman, M. 2000. *The Syntactic Process*. MIT Press.

Tellex, S.; Kollar, T.; Dickerson, S.; Walter, M.; Banerjee, A.; Teller, S.; and Roy, N. 2012. Approaching the symbol grounding problem with probabilistic graphical models. *AI magazine* 32(4).

Tenorth, M., and Beetz, M. 2012. A unified representation for reasoning about robot actions, processes, and their effects on objects. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

Tenorth, M.; Kunze, L.; Jain, D.; and Beetz, M. 2010. Knowrob-map - knowledge-linked semantic object maps. In *Humanoids*.

Topp, E. A.; Huettenrauch, H.; Christensen, H. I.; and Eklundh, K. S. 2006. Bringing together human and robotic environment representations-a pilot study. In *Int'l. Conf. on Intelligent Robots and Systems, 2006*, 4946–4952. IEEE.

Vasudevan, S., and Siegwart, R. 2008. Bayesian space conceptualization and place classification for semantic maps in mobile robotics. *Robotics and Autonomous Systems* 56(6).

Veloso, M.; Biswas, J.; Coltin, B.; Rosenthal, S.; Kollar, T.; Mericli, C.; Samadi, M.; Brandao, S.; and Ventura, R. 2012. Cobots: Collaborative robots servicing multi-floor buildings. In *Int'l. Conf. on Intelligent Robots and Systems (IROS), 2012*, 5446–5447. IEEE.

Wigdor, D., and Wixon, D. 2011. *Brave NUI world: designing natural user interfaces for touch and gesture*. Morgan Kaufmann.

Yang, J.; Yu, K.; Gong, Y.; and Huang, T. 2009. Linear spatial pyramid matching using sparse coding for image classification. In *Computer Vision and Pattern Recognition*, 1794–1801. IEEE.

Zender, H.; Mozos, O. M.; Jensfelt, P.; Kruijff, G.-J. M.; and Burgard, W. 2008. Conceptual spatial representations for indoor mobile robots. *Robotics and Autonomous Systems* 56(6).