A Robot Waiter that Predicts Events by High-level Scene Interpretation

Jos Lehmann[1], Bernd Neumann[1], Wilfried Bohlken[2] and Lothar Hotz[2]
[1]Cognitive Systems Laboratory, Department of Informatics, University of Hamburg, Hamburg, Germany
[2]HITeC e.V., University of Hamburg, Hamburg, Germany

Keywords: Prediction, Ontology, High-level Robotics.

Abstract: Being able to predict events and occurrences which may arise from a current situation is a desirable capability of an intelligent agent. In this paper, we show that a high-level scene interpretation system, implemented as part of a comprehensive robotic system in the RACE project, can also be used for prediction. This way, the robot can foresee possible developments of the environment and the effect they may have on its activities. As a guiding example, we consider a robot acting as a waiter in a restaurant and the task of predicting possible occurrences and courses of action, e.g. when serving a coffee to a guest. Our approach requires that the robot possesses conceptual knowledge about occurrences in the restaurant and its own activities, represented in the standardized ontology language OWL and augmented by constraints using SWRL. Conceptual knowledge may be acquired by conceptualizing experiences collected in the robot's memory. Predictions are generated by a model-construction process which seeks to explain evidence as parts of such conceptual knowledge, this way generating possible future developments. The experimental results show, among others, the prediction of possible obstacle situations and their effect on the robot actions and estimated execution times.

1 INTRODUCTION

The ability to look ahead and anticipate possible developments and events can be a valuable asset for robotic systems. By prediction, a service robot may provide timely assistance to elderly persons, anticipating their needs. A driver assistance system may brake when perceiving a rolling ball even before a child following the ball is visible. Robots seeking an obstacle-free path may anticipate the movements of persons crossing their way. There are several methodological approaches for realizing predictive power in robots, which will be discussed in the related-work section of this paper. Our approach is new in at least three respects. Firstly, it is based on an ontology with occurrence concepts which may be obtained by conceptualizing experiences. Secondly, predictions are performed by the same scene interpretation system which also recognizes occurrences actually happening in a scene. Thirdly, the knowledge representation framework connects high-level symbolic concepts with quantitative properties and elementary robot actions.

Our work is part of the project RACE (for Robustness by Autonomous Competence Enhancement) featuring a robot which learns from experiences. The RACE architecture, shown in Figure 1, integrates all essential robot functionalities around a common ontology and robot memory. Hence episodes experienced by the robot and instructions about how to perform a task can be used by the robot to establish new concepts and integrate these into the ontology. The concepts of the ontology are the basis for scene interpretation as well as prediction. Prediction is independent of the way concepts have been obtained, hence learning will not be addressed in this paper. The example domain of project RACE is a restaurant where the robot acts as a waiter. This is a highly dynamic domain with guests entering and moving about, persons or side tables occasionally blocking a path, and waiter activities ranging from serving guests to clearing tables. Hence predicting possible courses of events may be quite helpful.

The paper is structured as follows. Section 2 describes ontology-based scene interpretation as implemented in the framework SCENIOR (for SCEne Interpretation with Ontology-based Rules). Section 3 describes a running example of prediction performed by a robot in the restaurant domain. Section 4 describes the application of SCENIOR to the running

Lehmann J., Neumann B., Bohlken W. and Hotz L..
A Robot Waiter that Predicts Events by High-level Scene Interpretation.
DOI: 10.5220/0004819704690476
In *Proceedings of the 6th International Conference on Agents and Artificial Intelligence* (ICAART-2014), pages 469-476
ISBN: 978-989-758-015-4

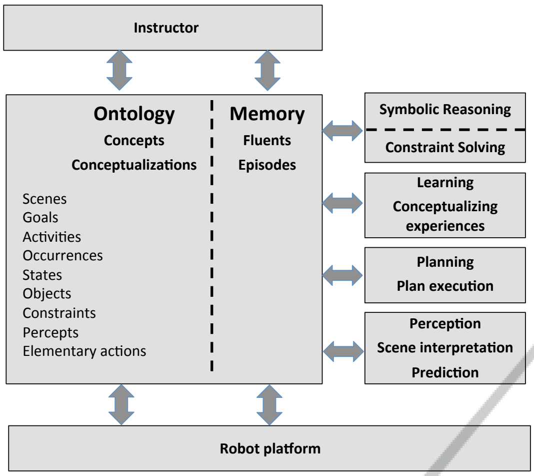

Figure 1: Relevant part of RACE architecture.

example. Section 5 evaluates the results. Section 6 discusses related work. Section 7 draws some conclusions.

2 ONTOLOGY-BASED SCENE INTERPRETATION

In this section we give an overview of the scene interpretation system SCENIOR which is integrated in the RACE system and used for scene interpretation as well as prediction. SCENIOR has been designed as a domain-independent framework for high-level scene interpretation. It can be adapted to different application domains by simply exchanging the conceptual knowledge base, represented in the standardized ontology representation language OWL[1] and augmented by constraints expressed in the semantic web rule language SWRL[2]. Figure 2 shows the main components of SCENIOR. The ontology can be used to automatically generate the knowledge structures and rules for an operational interpretation system, consisting of a JESS[3] rule engine, a constraint solver for quantitative temporal constraints, and an inference engine for probabilistic information in terms of Bayesian Compositional Hierarchies (BCHs) (Bohlken et al., 2013).

As described in (Neumann and Möller, 2006), conceptual structures for scene interpretation usually form compositional hierarchies consisting of aggregates at a higher abstraction level with aggregates at a lower abstraction level as parts, ‘properties’ in OWL syntax. In SCENIOR, compositional hierarchies of the ontology are converted into hypotheses structures which play the role of templates for the recognition process and for prediction. The tokens of a hypothesis structure represent the events which can be predicted.

The temporal structure of aggregates, specified by SWRL rules in the ontology, is converted into quantitative constraints on durations of components and in gaps between components on temporal relations between components in a temporal constraint net (TCN). Spatial information is represented in terms of events in predefined areas. The interpretation process is incremental and can operate in real-time for everyday dynamic scenes. Its input data are primitive states and occurrences as perceived by the robot’s perception system, and elementary robot actions logged by execution monitoring. As an example, a typical input could be (At guest1 doorArea 0:20:33 0:20:56), asserting that a guest is within a predefined door area in the given time interval.

The interpretation system, realized by the JESS rule engine, tries to assign evidence, obtained from low-level image analysis in terms of primitive states and occurrences, to leaves of the hypotheses structures, instantiating corresponding concepts. If there are several possibilities, the system establishes a separate interpretation thread for each alternative. The quantitative temporal information of incoming evidence is used to update the TCN. If the temporal constraints cannot be satisfied, the instantiation of that thread fails. When all parts of an aggregate are instantiated, the aggregate is instantiated as a whole and treated as input for higher-level aggregates. This way, a multi-thread interpretation process is realized, with fully instantiated hypotheses structures as final output.

Alternative interpretations can be ranked, also in intermediate interpretation stages, based on proba-

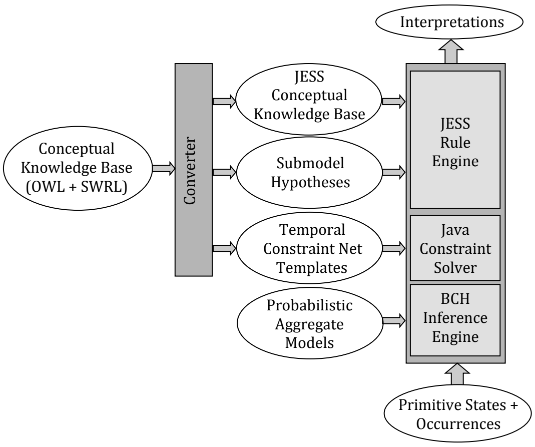
Interpretations
Conceptual Knowledge Base (OWL + SWRL)
Converter
JESS Conceptual Knowledge Base
Submodel Hypotheses
Temporal Constraint Net Templates
Probabilistic Aggregate Models
JESS Rule Engine
Java Constraint Solver
BCH Inference Engine
Primitive States + Occurrences

Figure 2: Components of SCENIOR scene interpretation system.

[1] http://www.w3.org/TR/owl2-overview/

[2] http://protege.cim3.net/cgi-bin/wiki.pl?SWRLLanguageFAQ

[3] http://www.jessrules.com/

bilistic aggregate models. This way, weak interpretation threads can be discarded, realizing a Beam Search. Probabilistic ranking is currently not used for prediction.

SCENIOR is designed to be robust against missing input, due for example to limitations of the robot's perception. To achieve that, SCENIOR has the ability to infer (to hallucinate in SCENIOR-jargon[4]) missing evidence if it helps to complete higher-level aggregates. This ability is also used to predict future developments of a scene, as will be described in Section 4.

3 EXAMPLE DEMONSTRATOR

Our guiding prediction examples deal with concepts which the robot has learnt in the scenarios described below. This is the short version of a longer demonstrator and it is meant to show how the robot predicts events. Note that in the following, the usual ontological naming conventions are used: all names of instance data (individuals) start with a lower case letter, comprise the name of their class (or an acronym thereof) and a integer at the end (except for the robot's name, which has no numerical index); names of concepts (classes) are compound and each component starts with a capital letter. All other references to individuals and classes are informal.

Figure 3 illustrates an experimental restaurant setting, which comprises: a counter (counter1), tables (e.g., table1, table2), people (e.g., guest1, sitting on chair), a coffee mug (e.g., mug1), a robot (trixi) and predefined reference areas for navigation (e.g. pre-manipulation and manipulation areas pmaSouth1 maSouth1) and manipulation (e.g. placing area paEast1).

The initial position of the robot is at the counter, i.e. in the area nearAreaCounter1 (which includes counter1's manipulation and pre-manipulation areas), where it has just picked up mug1 from counter1 and is ready to perform the task of serving it to guest1 at table1, approaching the guest from the right.

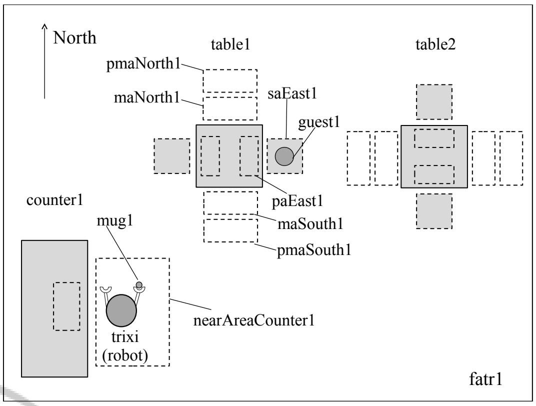

Figure 3: Floor plan of experimental restaurant setting.

Scenario A: The robot starts its navigation but finds table1's manipulation area north (maNorth1) blocked by a person (person1). The robot is instructed to wait until person1 has freed the path. After a short while, person1 frees the path and the robot completes its task.

Scenario B: The robot starts its navigation but an extension table exTable1 blocks maNorth1. Based on the experience in Scenario A, the robot decides to wait. After a while, it is instructed that this kind of obstacle must be circumnavigated, hence the robot chooses another path, thereby navigates to maSouth1 and completes its task.

Scenario C: Before starting the task anew and after having grasped mug1 from counter1, the robot is asked to predict, based on its previous experience, what may happen next. The robot will predict three possible alternative courses of events:

Course 1: maNorth1 will not be blocked, task will be completed.

Course 2: maNorth1 will be blocked by person, task will be completed as Scenario A.

Course 3: maNorth1 will be blocked by table, task will be completed as Scenario B.

4 ONTOLOGY-DRIVEN PREDICTION

We now describe ontology-driven prediction using the scene interpretation system SCENIOR. Prediction is realized as model construction, i.e. as a reasoning process which tries to explain evidence in terms of high-level structures and this way generates possible future evidence. We restrict prediction to partial model construction by considering only those conceptual structures which are compositionally connected to the given evidence. Prediction follows hypotheses structures in a similar way as a scene interpretation process, by constructing aggregate instantiations from components. Consider the general format of an aggregate:

[4] The term "hallucinate" reflects the idea that "perception is controlled hallucination" which in the Artificial Intelligence community is attributed to Max Clowes (1971), although various Internet sources date it as far back as the German physician and physicist von Helmholtz (1821-1894).

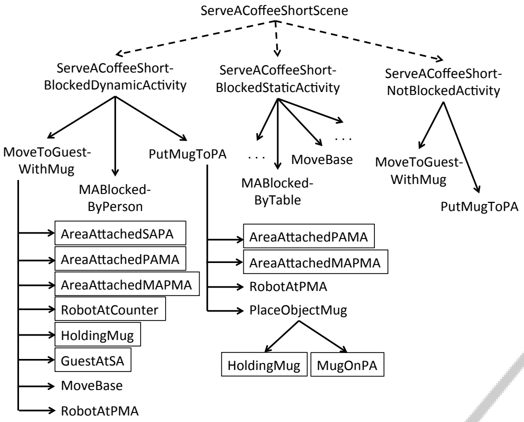

Figure 4: A multi-level compositional structure. Solid edges represent conjunctive properties, dotted edges disjunctive properties.

```
Class: <concept name>
  EquivalentTo / SubClassOf:
    <superconcept name>
  AND <property-1>
    <cardinality restriction-1>
    < property filler concept-1>
    ...
  AND <property-N>
    <cardinality restriction-N>
    < property filler concept-N>
```

The aggregate may be a property filler for a higher-level aggregate; simultaneously, properties of the aggregate may be connected to lower-level aggregates, see Figure 4 for an example of a multi-level compositional structure.

Given evidence for a property filler, model construction amounts to asserting the instance of the aggregate concept as a whole, and in consequence instances of all its other property filler concepts. Asserted instances are recursively treated as evidence, triggering further aggregate assertions. Since an asserted instance may be an aggregate with parts, the process may propagate top-down as well as bottom-up.

Instantiating a concept in the prediction process calls for a value assignment, and different values may lead to different alternative predictions, giving rise to a branching future. The following strategy is pursued:

1. Concepts with symbol values are assigned all possible instances of compatible class known so far and, under certain conditions, also a new instance.

2. Concepts with a numerical value range submit the current value range to the constraint system, leading to a reduced range or to inconsistency. This pertains, in particular, to all time intervals.

For the RACE domain, the basic idea is to let SCENIOR go ahead with the current scene interpretation irrespective of real-time, and hallucinate expected evidence, this way generating a prediction. To illustrate the process, consider again the compositional structure depicted in Figure 4. As described in the preceding section, the robot hat learnt to serve a coffee even if an obstacle is in the way. The figure shows the detailed compositional structure of the activities when the manipulation area is blocked by a person. The other two versions for a ServeACoffeeShort-Scene have the same structure except for differences in the middle level and in temporal constraints (not shown). The concept AreaAttachedSAPA specifies that a placing area (PA) is assigned to a sitting area (SA). Similarly, a manipulation area (MA) may be assigned to a PA, and a premanipulation area (PMA), where a robot prepares for a manipulation, may be assigned to a MA.

Note that by exchanging the ontology the prediction procedure is automatically adapted to a different domain.

We now consider the prediction task presented in the preceding Section in Scenario C. The robot has the goal to place the mug in front of the guest, it knows the area attachments as part of its permanent knowledge about the environment. The facts characterizing the situation are given in terms of instances of the corresponding concepts, also the goal which is part of the given prediction situation. All instantiated concepts are marked by boxes in Figure 4.

In real robot operations, evidence is provided by the robot's execution monitoring of its own activities, by the robot's observations of the environment, and by initialization with permanent knowledge. Evidence is presented as fluents using the YAML syntax, shown below for the instance guestAtSA1.

```
!Fluent
Class_Instance: [GuestAtSA, guestAtSA1]
StartTime: [00:00:00, 00:00:00]
FinishTime: [inf, inf]
Properties:
 -[hasPhysicalEntity, PhysicalEntity, guest1]
 -[hasArea, SA, saEast1]
```

The fluent specifies that the occurrence guestAtSA1, instance of class GuestAtSA, has begun at time 00:00:00 relative to the starting time of the episode, the finish time being unrestricted. The two bracketed time values can be used to denote an uncertainty range. The occurrence has two components, a guest guest1 and the predefined sitting area saEast1 of table1.

We now sketch the technical steps for ontology-based prediction with SCENIOR in this situation. As mentioned before, upon initialization SCENIOR creates hypotheses structures for all aggregate concepts of its ontology, including the ServeACoffee concepts depicted in Figure 4. Attached to the hypotheses structures are automatically generated interpretation rules, realized by the JESS rule system. The rules fire if evidence for any concept arrives. If the evidence fits several concepts, it is assigned to each alternative, and independent interpretation threads are created for the alternatives.

In our case, the evidence describing the prediction situation immediately causes the creation of six alternative threads representing possible courses of events, two for each of the three kinds of ServeACoffeeScene. For each kind, one of the two threads specifies area instantiations for a service from the north, the other for a service from the south. Since both components of PlaceObjectMug are introduced as evidence, the aggregate PlaceObjectMug is instantiated immediately, as a necessary robot activity to achieve goal mugOnPA postulated as evidence.

SCENIOR now performs prediction by "thinking ahead", realized by advancing a simulated time. At the beginning of the prediction phase, the temporal constraint nets in all threads of SCENIOR indicate that the robot should start moving (MoveBase) to the designated premanipulation area as a possible way to complete evidence for higher-level aggregates (and thus possibly achieve the goal). Hence MoveBase is hallucinated for each thread, i.e. instantiated in prediction mode without evidence. After a while (of simulated time), the robot reaches the designated premanipulation area, and the occurrence RobotAtPMA is hallucinated. In the threads where blocking is expected, this leads to a completed ServeACoffeeShortNotBlockedActivity since the PutMugToPA has been instantiated earlier.

For the other kinds of ServeACoffeeShortScene the hypotheses graphs imply that the manipulation area will be blocked and this can be observed by the robot. The occurrences MABlockedByPerson or MABlockedByTable are therefore hallucinated while the robot is approaching the premanipulation area. In the case of a person blocking the area, the robot has learnt to wait until the area will be freed, and then to continue serving the placement area from the anticipated manipulation area. In the case of a static obstacle, like a table blocking the manipulation area, the robot has learnt to turn around and move to the other side of the table, serving the guest from the left as an exception. These activities are hallucinated in their respective order as the simulated time advances, and finally the goal is achieved. The alternative threads allow to predict completion times based on the temporal model. As it turns out, they differ considerably for our slow robot waiter depending on the blocking situation.

Table 1: Expected minimal durations for serving a coffee.

Course of Activities	Start	Finish	Duration
(ServeACoffeShortScene)		(MugOnPA)	
NotBlockedAct.	14:48:28	14:49:13	00:00:44
BlockedDynamicAct.	14:48:28	14:49:43	00:01:15
BlockedStaticAct.	14:48:28	14:54:44	00:06:12

Note that SCENIOR typically entertains a large number of threads during a prediction process, often more than one hundred. The threads represent alternative partial predictions due to ambiguous assignments (several PMAs and MAs are possible) and also due to the strategy, adopted for real-life scene interpretation, to doubt all evidence. In our prediction experiments, the threads are rated by a measure of completeness, hence incomplete predictions are discarded at the end.

5 EXPERIMENTS AND EVALUATION

In this section, we describe experiments carried out with concrete predictions, and a first evaluation of the approach. The first prediction experiment is based on the ontological structures illustrated in Figure 4. SCENIOR has received background knowledge about area attachments (areaAttachedSAPA1, etc.), evidence about the current situation (guestAtSA1, robotAtCounter1, holdingMug1) and postulated evidence about the goal mugOnPA1.

Screenshots of alternative predictions determined by SCENIOR for this evidence are shown in Figures 5 and 6 for Course 2 and Course 3 of Scenario C, respectively, as described in Sections 3 and 4. The screenshot for Course 1 cannot be shown for lack of space. Downward arrows indicated the compositional structure of aggregates, upward arrows indicate instantiations. Evidence is depicted by white boxes (at the bottom), concepts instantiated through evidence by dark gray boxes (in the middle), and hallucinated instantiations by light gray boxes (in the top area). Each box also shows the ranges for the starting and finish time. For hallucinated instantiations, most ranges remain uncertain to some extent, according to the possible time intervals specified by the TCN.

For a real-life application, the expected minimal durations for serving a coffee shown in Table 1 would probably be the most interesting prediction data. As to be expected, the obstacle-free service takes the shortest time. Waiting for a person to move out of the way causes a slight delay. Turning around and travelling to the other side of the table when facing a static obstacle causes a major delay. In our experiment, the quantitative values result from durations de-

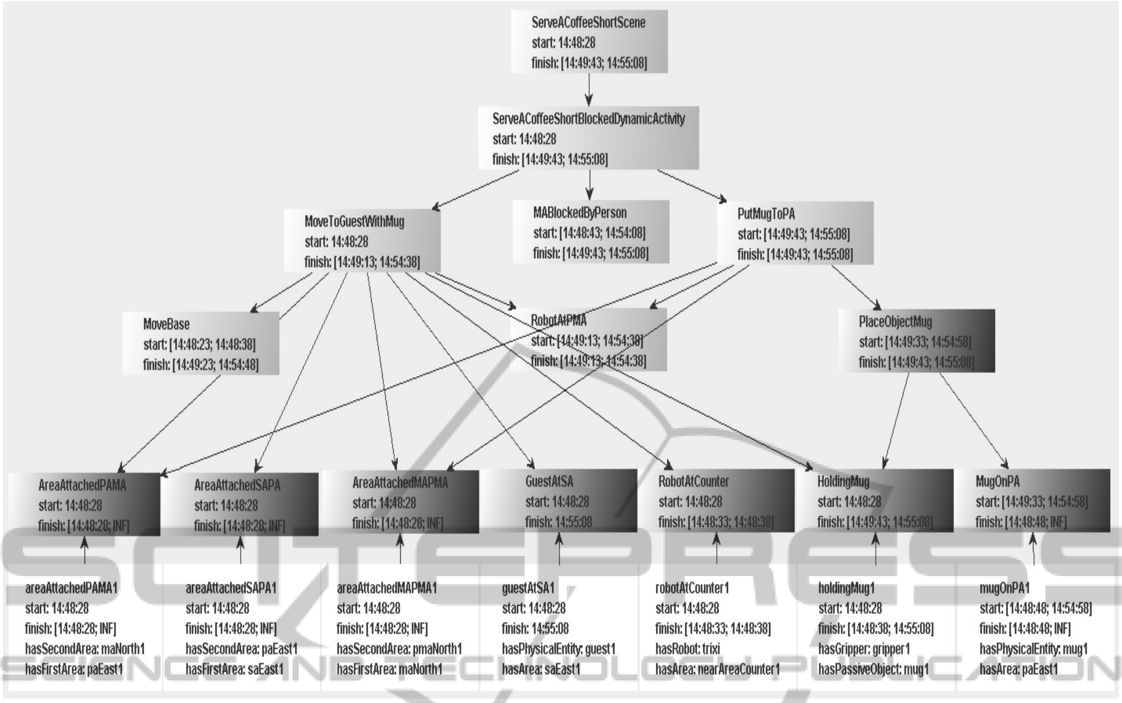

Figure 5: Prediction of occurrences (light gray) for ServeACoffeeShortBlockedDynamic after initial knowledge and goal (white and dark gray).

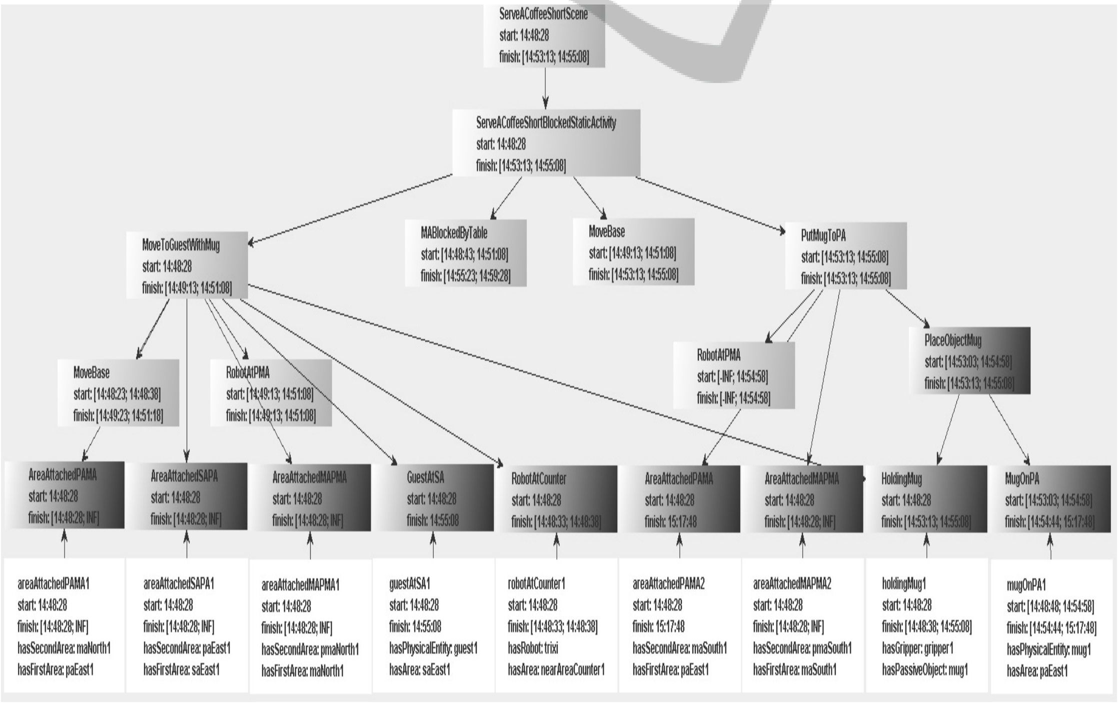

Figure 6: Prediction of occurrences (light gray) for ServeACoffeeShortBlockedStatic after initial knowledge and goal (white and dark gray).

fined in the ontology for each of the activity concepts, including the expected time for a person to unblock the way.

In total, SCENIOR has generated six complete alternative predictions for how the robot might achieve the goal, three as described above for attempting to serve from the north, and three very similar predictions for attempting to serve from the south. The computation time on a laptop has been 18s and 477 interpretation threads, all of which incomplete except for the six correct predictions. Whenever evidence enters the system, the number of existing interpretation threads doubles to reflect the possibility that the evidence may be faulty. Currently, this is applied to all evidence including background knowledge. In consequence, the number of threads often climbs above an upper limit, in the experiments set to 100, and is then reduced by discarding low-ranking threads. This strategy has been conceived for scene interpretation with noisy data, but it is also in some respect important in our prediction scenarios: The background knowledge provides two pieces of evidence for the concept AreaAttachedMAPMA, one referring to the areas north of the guest, the other to the areas south of the guest, only one of which will finally allow a complete interpretation. Hence at the time the evidence is provided, each must give rise to two alternative threads. We have shown in the preceding section that prediction is solely based on occurrence concepts represented in the ontology. By changing the ontology, predictions are immediately possible for a new domain. To illustrate this, we have employed prediction also for a second restaurant scene modelled as shown in Figure 7. Here a guest has entered at the door, and two developments of the scene are possible according to the model: (i) the guest may be a TransientGuest and leave without going to a table, or (ii) the guest may go to a table, have a coffee and complain (the reason is a late service).

Our experiments show that model-based scene interpretation can be used for prediction, with only minor changes to the interpretation system. Furthermore, the approach can be easily applied to other application domains, since the scene interpretation and prediction facilities are automatically generated from the ontological structures.

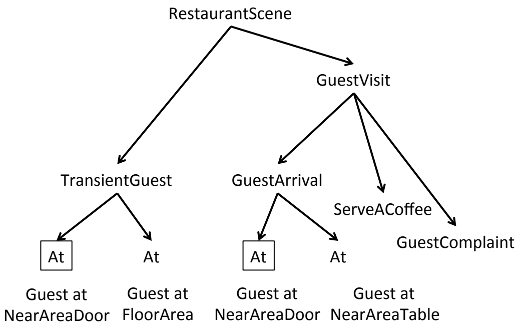

Figure 7: Model of guest visit.

6 RELATED WORK

In Robotics, prediction often refers to visual monitoring for obstacle avoidance. Given the role of the ontology in SCENIOR, our approach can better be compared to reasoning about action. A lot of the literature on ontology-driven prediction focusses on adapting for Description Logic (DL) the Action Calculi (ACs) developed between the 1960s and 90s. References to specific ACs can be found in (Thielscher, 2011). All ACs face core reasoning problems: the *Projection Problem* (how to compute the direct effects of an action); the *Ramification Problem* (how to compute the indirect effects of an action); the *Frame Problem* (how to compute what is not affected by the execution of an action). ACs are semi-decidable and as a consequence they can not be used by DL reasoners. Fragments have been identified to achieve automation (Baader et al., 2010) but these results do not easily scale up.

Other approaches to prediction are, like SCENIOR's (Bohlken et al., 2011), based on ontology-based scene interpretation. (Neumann and Möller, 2006) describe how evidence can be used to trigger model-based hypotheses about a scene which are used to predict parts not yet supported by evidence. The classical example is the observation of a ball running over a street, which can be taken as a partial instantiation of a model for a child chasing the ball.

A first formalization of scene interpretation based on model construction is owed to (Reiter and Mackworth, 1989). Here scene interpretation is a search for instantiations of the conceptual background knowledge such that the instantiations contain the evidence about the scene. A model constructed this way may naturally comprise predictions about the development of the scene. (Neumann and Möller, 2006) extends the model construction paradigm to ontologies using DL to represent knowledge. (Riboni and Bettini, 2012) check evidence for consistency with asserted interpretation, realizing model construction for fixed activities. (Cohn et al., 2003) and (Shanahan, 2005) formulate interpretation in terms of abduction, as the search for high-level concepts whose instantiation would entail the evidence. (Chen and Nugent, 2009) formulate interpretations as a two-tiered process of deriving an abstracted ontology from the data and of matching it with a standard ontology.

7 CONCLUSIONS

t has been shown that model-based scene interpretation can be used for prediction tasks. From a conceptual point of view, this is not surprising because both, prediction and scene interpretation, are model-construction tasks in the logical sense. For this reason, it is easy to see that the reasoning framework can also be used, besides for predicting, for reconstructing past occurrences, or generally, for imagining any kind of missing information, past and future, which serves to integrate given evidence into higher-level models.

As a draw-back, we must mention the tedious task of preparing hand-crafted models in OWL and constraints in SWRL. While this combination of symbolic reasoning and constraint solving is a promising architecture for bridging the gap between high-level concepts and low-level robot routines, standardized system support for incremental scene interpretation and prediction is not yet available, and a complex system like SCENIOR is required to operationalize real-life applications.

As work in RACE progresses, we expect that the robot will be able to learn models from experiences. This will hopefully allow to limit the production of hand-crafted models to basic behaviours and occurrences, from which higher-level aggregates can then be formed by learning.

Future work will also adapt an existing probabilistic rating system using Bayesian Compositional Hierarchies (BCHs) (Bohlken et al., 2013) for prediction tasks.

ACKNOWLEDGEMENTS

The research reported in this paper was partially supported by the European Grant FP7-ICT-2011-7-287752

REFERENCES

Baader, F., Lippmann, M., and Liu, H. (2010). Using causal relationships to deal with the ramification problem in action formalisms based on description logics. In Fermüller, C. G. and Voronkov, A., editors, *LPAR (Yogyakarta)*, volume 6397 of *Lecture Notes in Computer Science*, pages 82–96. Springer.

Bohlken, W., Koopmann, P., Hotz, L., and Neumann, B. (2013). Towards ontology-based realtime behaviour interpretation. In Guesgen, H. and Marsland, S., editors, *Human Behavior Recognition Technologies: Intelligent Applications for Monitoring and Security*, IGI Global, pages 33–64.

Bohlken, W., Koopmann, P., and Neumann, B. (2011). Scenior: Ontology–based interpretation of aircraft service activities – fbi–hh–b–297/11. Technical report, University of Hamburg, Department of Informatics Cognitive Systems Laboratory.

Chen, L. and Nugent, C. (2009). Ontology-based activity recognition in intelligent pervasive environments. *Int. J. Web Inf. Syst.*, 5(4):410–430.

Cohn, A. G., Magee, D. R., Galata, A., Hogg, D., and Hazarika, S. M. (2003). Towards an architecture for cognitive vision using qualitative spatio-temporal representations and abduction. In Freksa, C., Brauer, W., Habel, C., and Wender, K. F., editors, *Spatial Cognition*, volume 2685 of *Lecture Notes in Computer Science*, pages 232–248. Springer.

Neumann, B. and Möller, R. (2006). On scene interpretation with description logics. In *Cognitive Vision Systems*, pages 247–275.

Reiter, R. and Mackworth, A. K. (1989). A logical framework for depiction and image interpretation. *Artif. Intell.*, 41(2):125–155.

Riboni, D. and Bettini, C. (2012). Private context-aware recommendation of points of interest: An initial investigation. In *PerCom Workshops*, pages 584–589. IEEE.

Shanahan, M. (2005). Perception as abduction: Turning sensor data into meaningful representation. *Cognitive Science*, 29(1):103–134.

Thielscher, M. (2011). A unifying action calculus. *Artif. Intell.*, 175(1):120–141.