Situation-Specific Intention Recognition for Human-Robot Cooperation

Peter Krauthausen[*] and **Uwe D. Hanebeck**

Intelligent Sensor-Actuator-Systems Laboratory (ISAS),
Institute for Anthropomatics,
Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany.
Peter.Krauthausen@kit.edu, Uwe.Hanebeck@ieee.org

Abstract. Recognizing human intentions is part of the decision process in many technical devices. In order to achieve natural interaction, the required estimation quality and the used computation time need to be balanced. This becomes challenging, if the number of sensors is high and measurement systems are complex. In this paper, a model predictive approach to this problem based on online switching of small, situation-specific Dynamic Bayesian Networks is proposed. The contributions are an efficient modeling and inference of situations and a greedy model predictive switching algorithm maximizing the mutual information of predicted situations. The achievable accuracy and computational savings are demonstrated for a household scenario by using an extended range telepresence system.

1 Introduction

Recognizing a human's intentions, plans, and actions is crucial to the facilitation of non-verbal human-robot cooperation. Intention recognition [1] is the process of estimating the force driving human actions based on noisy observations of the human's interactions with his environment. For example, a robot embedded in a household can assist the human at its best, when estimating whether the human wants to *cook, wash, etc.*, based on observations such as his location, grasping activity, and object interactions [2]. In general, approaches to intention recognition may be categorized into symbolic approaches [3], probabilistic approaches, [4], [5] and blends thereof [6]. The key difference between symbolic and probabilistic approaches is that in the former the possibility of an intention is deduced, while in the latter the probability of an intention is inferred. In this paper, intention recognition is considered a discrete-time state estimation problem formalized in a *Dynamic Bayesian Networks* (DBN) [7], with a state containing the set of intentions to be estimated [2], [8], [9]. In order to allow for an intuitive interactive cooperation, efficient online inference in these models must be achieved. The main challenge to be addressed is the large size of the measurement systems caused by the large number of features sensed.

The problem of inference with large measurement models is at the intersection of three research areas: *structured Bayesian Networks* (BN) [10],[11], *switching models*,

[*] This work was partially supported by the German Research Foundation (DFG) within the Collaborative Research Center SFB 588 on "*Humanoid robots – Learning and Cooperating Multimodal Robots*".

and *sensor selection*. A hierarchical fusion of measurements is modeled by a measurement system in form of a BN. Existing approaches to modeling and solving large-scale *structured BNs* include *Object-Oriented BNs* (OOBN) [12] and *Situation-Specific BNs* [13]. Although facilitating the handling of large BN greatly by the introduction of reusable objects and class structures, it is hardly possible to construct a BN for a specific query at each time step and still obtain real-time performance. In contrast, *Multinets* [14] are focused on dynamically *switching* the structure of a single network based on the state estimate. Extending this approach to considering only a subset of the measurement model for one BN may be understood as a *sensor selection* problem [15] over time. The specific difficulty for the intention recognition lies in the complexity of the measurement system, i.e., layered information fusion. In [16], an approach to selecting subsets of sensors for such a model, maximizing the mutual information $\mathbf{I}$ between the state and sensor subsets, was proposed. For the used sensor synergy graph, $\mathbf{I}$ has to be computed for all pairs of sensors.

In this paper, inference in large models with many features is addressed, where number of features prohibits the exhaustive calculations in [16]. The specific properties of the intention recognition problem are exploited to circumvent the exhaustive calculation, by using the natural decomposition of the problem into situations. A model predictive switching between reduced models is proposed. The reduced models correspond to sets of intentions summarized in situations and include the respective subsets of the measurement system, i.e., BN with less nodes and smaller state spaces. Performing inference with the reduced models only, computational savings can be obtained at a modest loss in accuracy.

2 Intention Recognition

Given a specific situation, a human has a set of coarse intentions, which manifest in fine-grained interactions with the world—the actions. These definitions are abstract and need to be instantiated for each application at hand, i.e., the set of possible actions, levels of abstractions, and the intentions need to be found for each restricted task. Given these limitations, a temporal causal forward model (CFM) for specific tasks and situations is derivable, as depicted in Fig. 1. Nodes reflect quantities, e.g., intentions or actions, and edges correspond to relations among them. Intention recognition is the process of inferring the human state by fusing the various online incoming observations (e.g., video streams or tracking results) according to this CFM, which encodes the human's rationale. The inset in Fig. 1 shows a detailed view of a complicated measurement system. This causal forward model may be simply converted into a BN, i.e., a probabilistic graphical model, by interpreting each quantity, e.g., the intention, as a random variable and the relations among random variables as conditional densities, e.g., $f(\underline{a}_t|\underline{a}_{t-1},\underline{i}_t)$ These conditional densities f quantify how likely value combinations are. For our purpose, it is necessary to model relations between hybrid sets of variables, i.e., sets of continuous and discrete valued variables, as well as nonlinear dependencies. For this reason, hybrid BN with mixture conditional densities are employed [17], as they allow for a unified treatment. Fig. 2 shows a block diagram representation of the BN for intention recognition. Here, $\underline{d}_t$ denotes domain knowledge, $\underline{s}_t$ the situations, $\underline{i}_t$ the

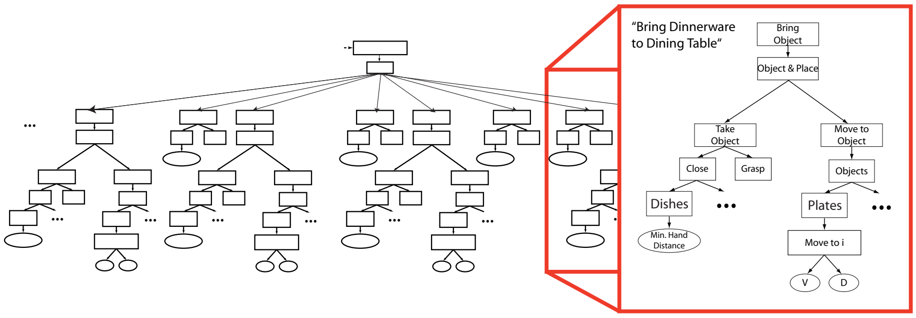

Fig. 1. Temporal causal forward model for one time step and detailed action recognition fragment, visualizing a part of the large measurement systems.

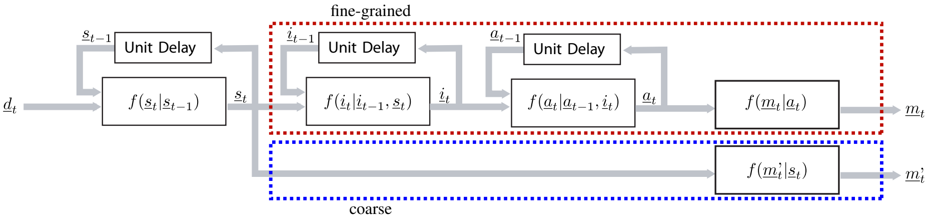
fine-grained
$\underline{s}_{t-1}$
Unit Delay
$\underline{i}_{t-1}$
Unit Delay
$\underline{a}_{t-1}$
Unit Delay
$\underline{d}_t$
$f(\underline{s}_t|\underline{s}_{t-1})$
$\underline{s}_t$
$f(\underline{i}_t|\underline{i}_{t-1}, \underline{s}_t)$
$\underline{i}_t$
$f(\underline{a}_t|\underline{a}_{t-1}, \underline{i}_t)$
$\underline{a}_t$
$f(\underline{m}_t|\underline{a}_t)$
$\underline{m}_t$
$f(\underline{m}'_t|\underline{s}_t)$
$\underline{m}'_t$
coarse

Fig. 2. Block diagram of the BN for the intention recognition. Note that the conditional density $f(\underline{m}_t|\underline{a}_t)$ subsumes the entire BN measurement system, e.g., corresponding to the inset in Fig. 1.

intentions, $\underline{a}_t$ the actions, $\underline{m}_t$ and $\underline{m}'_t$ the measurement variables for time step t, which will be explained in the next two sections. The corresponding conditional densities also are given, showing the temporal dependency of consecutive time steps. For the rest of this paper, it is assumed that $\underline{d}_t$ and $\underline{s}_t$ remain constant and only the model defined by $\mathscr{M}_t := \{\underline{i}_t, \underline{a}_t, \underline{m}_t\}$ will be considered. In order to infer the intention, information about the domain $\underline{d}_t$ and preceding time steps will be propagated forward, whereas measurements will be processed by a Bayesian backward inference step [17].

3 Online Model Switching

For non-trivial applications, the models constructed in Sec. 2 entail hierarchical and sequential models, relating atomic actions to multilevel action sequences. The resulting model is too large to allow for an interactive human-robot-cooperation. In this section, an online switching between reduced models is proposed to alleviate this problem.

Modeling Situations

A situation is a set of necessary conditions for human behavior, i.e., a set of conditions restricting the set of possible intentions. The constraints for situation s_i may be user given or determined automatically, e.g., spatio-temporal constraints may be obtained from an analysis of the human's movement profiles. The relations are assumed to be given in conjunctive normal form, similar to

$$\text{Cooking} : \text{'Near stove'} \wedge (\text{'12 AM'} \vee \text{'Food present'}) \equiv s_i : (\wedge \vee \underline{m}'_{i,t}) . \qquad (1)$$

The $m'_{i,t}$ correspond to measurable quantities. Note that $\underline{m}'_t = [m'_{d,t} \dots m'_{e,t}]^{\mathrm{T}}$ is not identical with $\underline{m}_t$. Typically, $m'_{i,t}$ are coarse features used exclusively for situation assessment. Incorporating these constraints into the overall BN corresponds to appending a variable $\underline{s}_t = [s_{1,t} \cdots s_{n,t}]^{\mathrm{T}}$, where each entry corresponds to a specific situation. For each s_i, variables for the terms The $m'_{i,t}$ are introduced and connected via *inverse-OR* nodes. The resulting clauses are then connected to *inverse-AND* nodes to arrive at a BN corresponding to (1). Performing this operation for all $s_{i,t}$ yields a measurement system for the situation model. In Fig. 2, this measurement system corresponds to the bottom system. Yet, as the situation is causal to the intentions $\underline{i}_t$ the situation BN can be merged with the intention recognition BN by introducing the dependency $f(\underline{i}_t|\underline{s}_t)$. One obtains a larger BN, where the intention recognition works as a fine-grained measurement model and the newly constructed situation BN functions as a coarse model. The dynamical development of the situation is accounted for by introducing $f(\underline{s}_{t+1}|\underline{s}_t)$.

Reduced Models

The complete BN for the intention recognition $\mathscr{M}$ resembles the causal forward model in Fig. 1 and is substituted by a set of reduced models $\mathscr{M}^k$

$$\mathscr{M}_t := \{\underline{i}_t, \underline{a}_t, \underline{m}_t\} \equiv \left\{\mathscr{M}_t^1 := \left\{\underline{i}^1_t, \underline{a}^1_t, \underline{m}^1_t\right\}, \mathscr{M}_t^2 := \left\{\underline{i}^2_t, \underline{a}^2_t, \underline{m}^2_t\right\}, \dots\right\} . \qquad (2)$$

The model $\mathscr{M}_t$ is defined by all intentions $\underline{i}_t = [i_{1,t} \cdots i_{o,t}]^{\mathrm{T}}, i_{j,t} \in \mathscr{A}_t$, all action subsystems $\underline{a}_t$ and the respective measured variables $\underline{m}_t$ as well as the respective probabilistic models, e.g., $f(\underline{i}_t|\underline{s}_t)$. The reduced models $\mathscr{M}_t^k$ are defined by only some intentions $\underline{i}_t^k = [i_{1,t} \cdots i_{h,t}]^{\mathrm{T}}, i_{j,t} \in \mathscr{A}_t^k \subseteq \mathscr{A}_t$ and the respective action subsystems with corresponding measured variables defined accordingly. The index t emphasizes the dependency on the time step. Note that the conditional densities need to be adapted for switching between different state spaces, e.g., over time $f(\underline{i}_{t+1}^k|\underline{i}_t^k)$ and $f(\underline{i}_t^k|\underline{s}_t)$. The coarse static measurement system of $\underline{m}'_t$ for the situation recognition is contained in all $\mathscr{M}_t^k$, to allow for fast situation recognition. This measurement system is not subject to switching.

Switching Algorithm

Based on the above situation model and the $\mathscr{M}_t^k$, an online model switching algorithm is proposed for choosing the model $\mathscr{M}_{t+1}^k$, which maximizes the reduction of uncertainty, measured by the mutual information $\mathbf{I}$, over the future situation $\underline{s}_{t+1}$ given $\underline{i}_{t+1}^k$. Additionally, a term $P(\mathscr{M}_t, \mathscr{M}_{t+1})$ penalizing a frequent model switching is added to the objective function

$$\mathscr{M}_{t+1}^* = \arg\max_{\mathscr{M}_{t+1}^k} \underbrace{\mathbf{I}(\underline{s}_{t+1}; \underline{i}_{t+1}^k) + \lambda\, P(\mathscr{M}_t, \mathscr{M}_{t+1}^k)}_{=:V(\mathscr{M}_{t+1}^k)} . \qquad (3)$$

Maximizing (3), the model $\mathscr{M}_{t+1}^*$ is selected, which is most consistent with the expected situation at $t+1$, which is predicted using $f(\underline{s}_{t+1}|\underline{s}_t)$. The parameter λ is a weight, representing our belief in the need for penalization. Furthermore, is assumed, that the $\mathscr{M}_t^k$ are discrete-valued only.

Alg. 1 summarizes the online model selection. It should be noted that this algorithm is greedy and only selects the best model with respect to (3) for a one-step horizon. Additionally, in contrast to [16], not $\mathbf{I}(\underline{s}_{t+1};\underline{m}_{t+1})$ is calculated, but an approximation in form of the intention is calculated—allowing an online application of the approach. Additionally, the calculation of $\mathbf{I}(\underline{s}_{t+1};\underline{i}_{t+1}^k)$ is approximated by neglecting the dependency $f(\underline{i}_{t+1}|\underline{i}_t)$. Besides the dynamics of $\underline{s}_{t+1}$ and $\underline{i}_{t+1}$, the measurement systems might contain dynamic dependencies too, which need to be considered. Regarding scalability, the algorithm is based on the decomposability of the human's intentions according to situations. The performance regarding estimation quality and computational savings is governed by the specific decomposition chosen. The experiments show, that given such a decomposition, the approach allows for efficient intention recognition.

Algorithm 1 Situation and Intention with Online Model selection

```
Input: Models $\mathscr{M}_t^k$, initial model $\mathscr{M}_0$, $\lambda$, penalty function $P(.,.)$
  for all Time-steps $t$ do
    Perform inference with $\mathscr{M}_t^k$, e.g., message passing [10], [17] $\rightarrow f(\underline{s}_t)$ // Inference
    for all $\mathscr{M}_{t+1}^k$ do
      Calculate $G_t^k := V(\mathscr{M}_{t+1}^k)$ // Model evaluation
    end for
    ${\mathscr{M}_{t+1}^k}^* = \arg\max_{\mathscr{M}_t^k} G_t^k$ // Model selection
  end for
```

4 Experiments

The proposed approach will be implemented for close cooperation with a humanoid robot in a household setting. To validate the approach, an extended range telepresent virtual household scenario is employed. In the rest of this section, this testbed and the results are discussed.

Telepresent Virtual Household Setting For the experiments, the *extended range telepresence system* from [18] was used. It allows the user to move in a virtual 1:1-scale kitchen model of the original household. Head and hand positions are tracked by an acoustic system in his real environment and are mapped to the virtual world. Thus, the human test person moves naturally in his environment and the virtual environment. The noise characteristics are similar to the expected capability of the vision system of the real robot. Additionally, the human grasping activity is measured by means of a bluetooth cyber-glove device. During the actual experiments, a test person was instructed to carry out a typical action sequence in a kitchen: *lay table*, *prepare a meal* followed by *clean dishes*. For this sequence, the test person has to cross a room, pick up some dinner ware, and bring it to the table on the opposite side of the kitchen. After fetching some ingredients, these are put into cooking pot. The pot is put onto the stove. Later, the initially used dinnerware is picked up and put into the dishwasher. For this test setup, only a subset of the actually available situations and intentions is used.

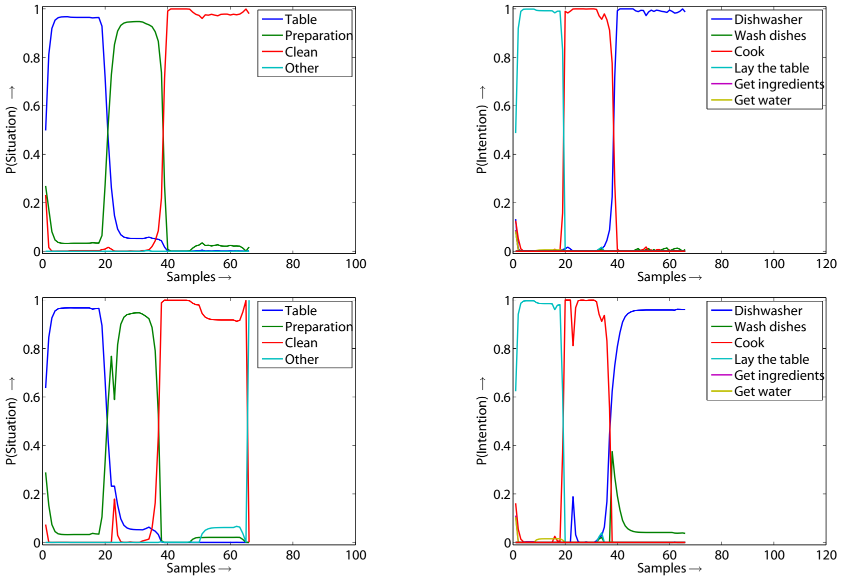

Fig. 3. Situation (left) and intention estimates (right) using the full (top) and reduced (bottom) models with the proposed online switching approach over time.

Results In this section, the achieved accuracy is compared to the computational savings of using the reduced models only for two model sizes. In Ex. 1 ten intentions were estimated with a BN of 319 nodes. Ex. 2 comprises 15 intentions and 625 nodes. Each model was decomposed into four reduced models with about 200-300 nodes per reduced model. For the small model, the situation and intention estimates are given in Fig. 3. The top images in Fig. 3 show the estimation results over time for the unswitched approach and the lower images show the estimation results for the switched approach. The estimates of the two approaches align nicely. Yet, some abrupt changes in the estimates can be observed. These occur when the model switches, as can be seen in Fig. 4, due to the different parametrization and state spaces of the models. In Fig. 4, the error, as the absolute difference between the estimates using the full and the reduced model, are given. Tab. 1 gives a statistic of the error in the situation and intention estimates for both models. In these tables the maximum and average mean absolute difference over all situations/intentions is given with its variance. The errors are averaged over time. Throughout the experiment the average error is modest, except where the model is switched. This leads to the conclusion that the approach works well, but a smooth model transition needs to be further investigated. The approach is relatively robust to misclassifications as can be seen in Fig. 4 for Ex. 2 where a situation is misclassified, leading to a drastic increase in error but after an additional situation change the approach aligns with the true situation. Regarding the computational savings, we compare the average number of nodes used in the BN given in Tab. 1 for Exp. 1 and Exp. 2. In both experiments, the employed BN are structured like Fig. 2. For both experiments the average number of nodes used is significantly reduced and the computation time is decreased by up to one order of magnitude. Note, even though Exp. 2 contains larger models, only a modest increase in the average number of nodes can be observed. In Exp. 2 larger parts of the model may be ignored, showing that the result will scale favorably for real scenarios.

5 Conclusion

In this paper, an approach to efficient intention recognition based on situation-specific model switching is proposed. To this end, a way of modeling situations and its integration with a BN, encoding the causal forward model of the human's rationale, was explicated. Based on a measurement system for the coarse situation assessment, which checks constraints for objects and places, a model predictive approach to online switching of reduced Dynamic Bayesian Networks for the intention recognition was proposed. At each time step, the best reduced model is selected based on the maximization of mutual information of the predicted situation and intention estimates, avoiding the expensive mutual information calculations for all sensors. Using an extended range telepresence system, the proposed approach was shown to deliver computational savings of up to one order of magnitude at an acceptable error level. In future work, it should be investigated if longer prediction horizons allow for improved switching performance and the neglected dynamic dependencies may be considered.

Table 1. Maximum and average difference between the situation estimates $\underline{s}_t$ (intention estimates $\underline{i}_t$) using the full and the reduced model and the respective standard deviation for both experiments. On the right, the used average number of nodes and computation time per step are given.

	$\underline{s}_t$	Absolute Error	$\underline{i}_t$	Absolute Error		Full	Switched
Ex. 1:	Max.	0.050 ±0.0099	Max.	0.060 ±0.0112	Avg. #Nodes	319	224.7
	Avg.	**0.025 ± 0.0024**	Avg.	**0.012 ± 0.0004**	Avg. Time in s	0.032	0.025

	$\underline{s}_t$	Absolute Error	$\underline{i}_t$	Absolute Error		Full	Switched
Ex. 2:	Max.	0.274 ±0.1701	Max.	0.325 ±0.1584	Avg. #Nodes	625	233.1
	Avg.	0.137 ±0.0428	Avg.	0.065 ±0.0063	Avg. Time in s	**0.234**	**0.026**

References

1. Schmidt, C.F., Sridharan, N.S., Goodson, J.L.: The Plan Recognition Problem: An Intersection of Psychology and Artificial Intelligence. Artificial Intelligence **11**(1–2) (1978) 45–83
2. Schrempf, O.C., Hanebeck, U.D., Schmid., A.J., Wörn, H.: A Novel Approach to Proactive Human-Robot Cooperation. In: Proceedings of the 2005 IEEE International Workshop on Robot and Human Interactive Communication (ROMAN 2005), Nashville, Tennessee (2005) 555–560

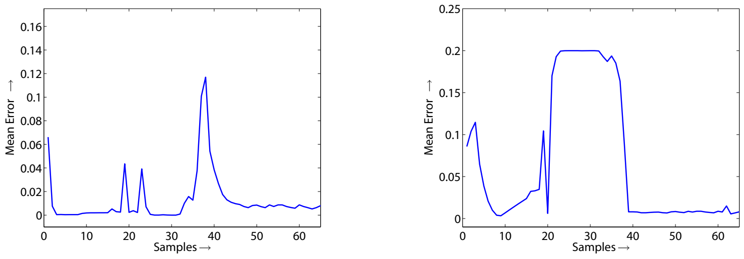

Fig. 4. Mean absolute error of $\underline{i}_t$ over time for Ex. 1 (left) and Ex. 2 (right).

3. Kautz, H.A.: A Formal Theory of Plan Recognition and its Implementation. In: Reasoning About Plans. Morgan Kaufmann Publishers, San Mateo, California (1991) 69–125
4. Pynadath, D.V., Wellman, M.P.: Probabilistic State-Dependent Grammars for Plan Recognition. In: Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI-00), San Francisco, California, Morgan Kaufmann Publishers, INC. (2000) 507–514
5. Bui, H.H.: A General Model for Online Probabilistic Plan Recognition. In: Proc. of the International Joint Conference on Artificial Intelligence (IJCAI). (2003) 1309–1315
6. Geib, C.W., Goldman, R.P.: Partial Observability and Probabilistic Plan/Goal Recognition-ijcai-05 workshop on modeling others from observations (2005)
7. Murphy, K.: Dynamic Bayesian Network : Representation, Inference and Learning. PhD thesis, UC Berkeley (2002)
8. Tahboub, K.A.: Intelligent Human-Machine Interaction Based on Dynamic Bayesian Networks Probabilistic Intention Recognition. Journal of Intelligent and Robotic Systems 45(1), 31-52. (2006)
9. Krauthausen, P., Hanebeck, U.D.: Intention Recognition for Partial-Order Plans Using Dynamic Bayesian Networks. In: Proceedings of the 12th International Conference on Information Fusion (Fusion 2009), Seattle, Washington (2009)
10. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan-Kaufmann Publishers, INC. (1988)
11. Koller, D.: Probabilistic Graphical Models : Principles and Techniques. MIT Press, Cambridge, Mass. [u.a.] (2009)
12. Koller, D., Pfeffer, A.: Object-Oriented Bayesian Networks. In: Proceedings of the 13th Annual Conference on Uncertainty in AI (UAI 1997), Providence, Rhode Island (1997) 302–313
13. Laskey, K., Mahoney, S.: Network Fragments: Representing Knowledge for Constructing Probabilistic Models. In: Proceedings of the 13th Annual Conference on Uncertainty in AI (UAI 1997), Providence, Rhode Island (1997) 334–341
14. Bilmes, J.: Dynamic Bayesian Multinets. In: Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI), San Francisco, California, Morgan Kaufmann Publishers, INC. (2000) 38–45
15. Williams, J.L.: Information Theoretic Sensor Management. PhD thesis, Massachusetts Institute of Technology (2007)
16. Zhang, Y., Ji, Q.: Efficient Sensor Selection for Active Information Fusion. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics **pp**(99) (2009) 1083–4419
17. Schrempf, O., Hanebeck, U.D.: Evaluation of Hybrid Bayesian Networks using Analytical Density Representations. In: Proc. of the 16th IFAC World Congress (IFAC 2005), Czech Republic (2005)
18. Rößler, P., Beutler, F., Hanebeck, U.D., Nitzsche, N.: Motion Compression Applied to Guidance of a Mobile Teleoperator. In: Proceedings of the 2005 IEEE International Conference on Intelligent Robots and Systems (IROS 2005). (2005) 2495–2500