Combined Intention, Activity, and Motion Recognition for a Humanoid Household Robot

Dirk Gehrig, Peter Krauthausen, Lukas Rybok, Hildegard Kuehne,
Uwe D. Hanebeck, Tanja Schultz, and Rainer Stiefelhagen
{dirk.gehrig, peter.krauthausen, lukas.rybok}@kit.edu

Abstract— **In this paper, a multi-level approach to intention, activity, and motion recognition for a humanoid robot is proposed. Our system processes images from a monocular camera and combines this information with domain knowledge. The recognition works on-line and in real-time, it is independent of the test person, but limited to predefined view-points. Main contributions of this paper are the extensible, multi-level modeling of the robot's vision system, the efficient activity and motion recognition, and the asynchronous information fusion based on generic processing of mid-level recognition results. The complementarity of the activity and motion recognition renders the approach robust against misclassifications. Experimental results on a real-world data set of complex kitchen tasks, e.g., *Prepare Cereals* or *Lay Table*, prove the performance and robustness of the multi-level recognition approach.**

I. INTRODUCTION

Humanoid robots are specifically aimed at supporting humans in every-day life tasks. In order to support the humans at their best, humanoid robots need to behave interactively like humans. This paper addresses video-based human behavior recognition. The recognition needs to be performed on-line and in real-time in order to allow the robot to react quickly to human behavior, i.e. intentions, activities, and motions. As the estimates are input to the control loop of the humanoid, the estimation quality and robustness needs to be high, as it directly impacts the robot's usability.

Modelling the behavior of the human in terms of intentions causing manipulations of the world, which may be modeled coarsely as activities and more fine-grained as sequences of motions, cf. Fig. 1, corresponds to modelling the causal dependencies of the human's rationale. For example, the task *Prepare Cereals* may be coarsely described as movements and manipulations in a specific area of a kitchen. We term this an *activity*. In contrast, the motion sequences of the task can be modeled in detail as a sequence of clearly defined *motion primitives*, such as *Place Object on Table*, *Pour*, or *Stir*. An *intention* combines these models with domain knowledge, e.g., object presence or time of the day. The intention as obtained from higher-level dynamics, domain knowledge, and various lower-level estimates is modeled by a Hybrid Dynamic Bayesian Network (HDBN) [1]. The data-driven discriminative modeling of the activities is performed by Support Vector Machines (SVM) [2]. For the detailed modeling of motion sequences, a set of motion primitives is used. Motion primitives are modeled as Hidden Markov Models (HMM) [3] and serve as an alphabet for a context-free grammar describing motion sequences. Our multi-level approach integrates the different levels of modeling the human behavior.

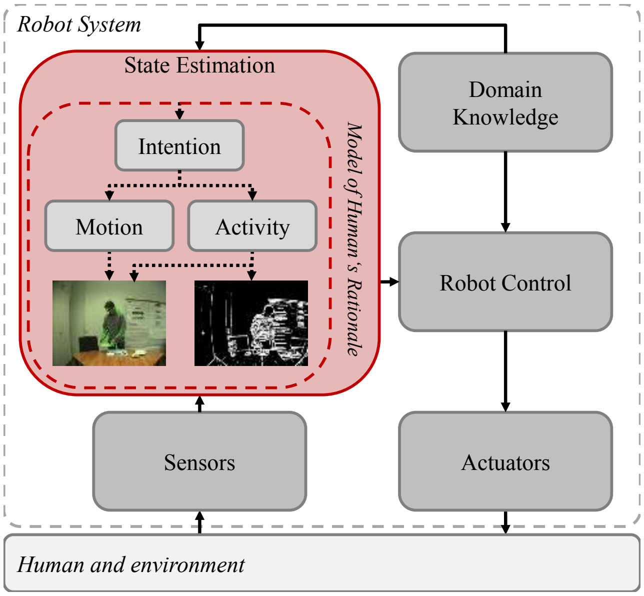


Fig. 1. Control loop of a robot: The combined intention, activity, and motion recognition is used for estimating the state of the human. This state estimate together with domain knowledge is input to a higher level control system governing the robot's actuators, i.e., his manipulations of the world.

II. RELATED WORK

The relevant literature is grouped according to the individual parts of the multi-level approach. *Intention recognition* is the inference of the force driving a human's behavior [4] based on observation of his manipulations. The existing approaches may be categorized according to the consideration of uncertainty induced by sensor noise and temporal unobservability. Symbolic approaches, often generalizing to

D. Gehrig and T. Schultz are with the Cognitive Systems Lab (CSL), P. Krauthausen and U. D. Hanebeck are with the Intelligent Sensor-Actuator-Systems Laboratory (ISAS), and L. Rybok, H. Kuehne, and Rainer Stiefelhagen are with the Computer Vision for Human-Computer Interaction Lab (CV:HCI), Institute for Anthropomatics, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany. R. Stiefelhagen is also with the Fraunhofer IOSB, Karlsruhe, Germany. 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 by the German Federal Ministry of Education and Research as part of the GEMS program.

plan recognition, have been successfully employed, e.g., in software agents [5]. Probabilistic intention recognition with probabilistic graphical models has been developed in the field of security and surveillance [6]. In robotics, probabilistic intention recognition has been employed amongst other applications for wheel-chair steering support [7]. The work resembling this paper most in terms of models and inference methods is [4]. Our work presented here extends this approach to achieve higher robustness and incorporate asynchronous mid-level measurements.

Activities denote complex motion sequences, which get their meaning from the overall situation context. A detailed overview of the current state-of-the-art in *activity recognition* can be found in, e.g., [8], [9]. Typical approaches model dependencies between simple actions with graphical models [10], grammars [11], or knowledge bases [12]. Since activities in a household scenario usually consist of a quasi-periodic repetition of short action sequences, we follow a different strategy and infer the activity within a temporal window directly from video features. Our approach is motivated by works in the field of space-time interest points based recognition of basic actions [13]. However, in contrast to other works, it can be and has already been successfully applied to complex real-world scenarios.

Motion recognition is the recognition of fine-grained motion primitives, which are part of more complex human motion sequences. Modeling motion sequences with primitives has for example been used in imitation learning and programming by demonstration [14]. A well-known statistical approach to primitive modeling are HMMs [3], which are suitable for modeling the sequential nature of motion primitives [15]. For the combination of motion primitives to longer motion sequences, context-free grammars have been proposed [11]. We extended the motion recognition system in [16] to an on-line recognition system and by learning our grammar automatically. The system returns the recognized motion primitives during the performed motions in real-time.

III. Multi-level Approach

The multi-level approach presented in this paper corresponds to the estimator for the state of the human in the overall robot control loop shown in Fig. 1. The estimator consists of the combined intention, activity, and motion recognition, i.e., the intention recognition integrates the activity and motion estimates as well as uncertain domain knowledge. Two key ideas govern this approach: *modularity* and *consistent uncertainty treatment*. The proposed system consists of *modular* representations, which are trained separately. Inferring the current state integrates information from all components. The *consistent processing of uncertain information* corresponds to a propagation of uncertainties about estimates from all components through the overall system. This consistent processing is required to allow for robust stochastic control, e.g., robustness against light-dependent image noise. Information passing at the systems' interfaces thus corresponds to exchanging posterior probability distributions. The modularity and consistent uncertainty treatment allow for an easy extension of the approach to include more classifiers. Every stand-alone classifier module which outputs posterior probability densities can easily be added as a subsystem. In the rest of this section, all parts of the entire system are described - from the used low-level features to the intention recognition.

A. Low-level visual features

The features used by the vision-based modules encode motion and appearance. Treating both feature types independently is, according to recent studies in neuro-science [17], in line with the way humans perceive movement in their environment.

a) Histogram of Sparse Flow: The motion features are based on histograms of global sparse optical flow obtained from feature tracking, representing every frame of the image sequence by a global histogram of its overall motion directions without any further local information. The weighted histogram[1] $H_t^f =: \underline{\hat{v}}'_t$ for frame t is calculated from the motion vector of the feature points of images I at time index t and $t+1$ (I_t, I_{t+1}). The motion vector $(u(\delta t), v(\delta t))$ of the feature is used to calculate the resulting motion direction θ (an angle value from $[-\pi, \pi]$) and γ defining the motion intensity. The elements for one bin of the histogram are calculated based on the motion angle θ. The bin entries are weighted with the respective motion intensity γ. The k-th bin of the weighted histogram is calculated from the intensity of all elements with the related motion direction.

b) Histogram of Oriented Gradients: Analogous to the motion features, the appearance of a scene is encoded using weighted histograms of dense image gradients. At time t, each pixel of a gradient map contributes to the bin of a histogram H_t^g which corresponds to the pixel's discretized orientation angle. Each histogram contribution is weighted by the gradient's magnitude in order to lower the effect of noise. In our experiments, we set the histogram size to 30 bins for both feature types.

B. Motion Recognition

The motion recognition uses the low-level motion features $\underline{\hat{v}}'_t$ to recognize the motions of an observed person. The motions are modeled as a concatenation of motion primitives such as *Place Object on Table*. This allows a very flexible and robust modeling of a large variety of motion sequences. Each motion primitive is represented by an HMM, which models the sequential nature of the motion primitive and can be optimized incrementally. The possible concatenations of the motion primitives are modeled using an automatically learned context-free grammar. The following paragraphs describe the components of our motion recognition system, i.e., the input features, the model topology, the model initialization, training, and optimization, as well as the decoding strategy. The system applies the one pass IBIS decoder [18], which is part of the Janus Recognition Toolkit JRTk [19].

[1] The time is indexed by t, sequences are denoted by $t : 0$ and observed values by $\hat{y}$. Random variables are printed bold and vectors are underlined.

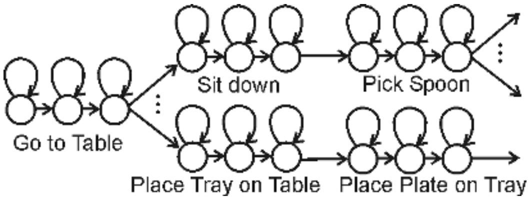


Fig. 2. Sequence of HMMs modeling flexible human motion sequences.

Features: As visual features for motion recognition, the global sparse optical flow histograms $\underline{\hat{v}}'_t$ presented in Sec. III-A are used. The histograms are sampled over time, resulting in 30-dimensional input vectors for the HMMs. For a good recognition rate the input vectors are normalized over time.

Representation: Each motion primitive is modeled with a linear left-to-right HMM. Each state of the left-to-right HMM has two equally likely transitions, one to the current state and one to the next state. The emission probabilities of the HMM states are modeled by Gaussian mixtures. The number of states and the number of Gaussians per mixture were optimized in the cross-validation experiments described below. A motion sequence is modeled as a sequential concatenation of these motion primitive models (see Fig. 2) using an automatically learned context-free grammar. We extended the Sequitur algorithm [20] to work on a set of motion sequences instead of only one sequence.

Learning: To initialize the HMM models of the motion primitives, we manually segmented the histogram sequences into the motion primitives. The manually segmented data are then equally divided into N sections for each motion primitive, where N is the number of states in the HMM. A Neural Gas algorithm [21] is applied to initialize the HMM-state emission probabilities for each state. We then perform ML-training on the unsegmented motion sequences using the Viterbi algorithm.

Recognition: Decoding of the system is carried out by a time-synchronous beam search. The most likely motion sequences $\underline{m}^*_{t:0}$ (sequences of motion primitives) are calculated based on the feature sequence $\underline{\hat{v}}'_{t:0}$. Large beams are applied to avoid pruning errors. To guide the recognition process, we use the automatically generated context-free grammar. Our automatically created grammar performs slightly better than a manually created one and is built a lot faster. The grammar allows a flexible and reliable recognition of the possible motion sequences. The input features are processed once every second and those motion sequences are calculated that match the data best up to the current time step. The log-likelihoods of the latest motion primitives (one per sequence) are normalized and passed to the intention recognition system as an approximation of the posterior probability distribution.

C. Activity Recognition

The activity recognition gives a coarse (in the sense of temporal resolution), but accurate estimation about the situation inside a room with a high update rate. Our approach is based on the *bag-of-words* method that already has been successfully applied to the problems of classifying objects [22] and basic actions [23].

Feature representation: It has been shown in [24], that only a few frames suffice to discriminate unambiguous basic actions. Our feature representation extends this principle to the recognition of activities by exploiting the nature of common household tasks, which mainly consist of a quasi-periodical repetition of short motion sequences. For instance, the activity *Lay Table* may consist of a repeated execution of the motion *Pick up Object* followed by *Place Object on Table*. Since such motion sequences define an activity, we reason that it is sufficient to base the recognition on activity snippets that last at least as long as one motion sequence period. Thus, we apply a sliding window to the input image sequence in order to obtain successive activity estimates.

Within each temporal window, we identify spatio-temporal regions of interest (ROI) in which the low-level motion and appearance histogram features H_t^f and H_t^g are calculated for each frame. The location and spatial size of the gradient ROI is determined by employing a fast 2D interest point detector [25] to every fourth frame. Since our optical flow field is very sparse, we calculate only one optical flow ROI per frame based on the difference of successive images. The temporal extension of the ROI for both feature types is fixed to a duration of 10 frames. All frame-based low-level histogram features within a ROI are further accumulated and normalized to form a spatio-temporal cuboid histogram feature. Regarding spatial and temporal dimension in such an independent way makes feature calculation very fast and combines the advantages of space-time interest points and dense feature sampling.

Finally, we combine all cuboid features within a temporal window with two bag-of-words models [22], one for each cuboid feature type. In order to reduce quantization errors when calculating the bag-of-words histograms, we employ a soft-voting scheme as described in [26]. The resulting histograms are then concatenated to one vector $\underline{\hat{v}}_t$, which is used to infer the activity.

Learning and Recognition: We map the features $\underline{\hat{v}}_t$ to one of the activity classes using an SVM with an RBF-kernel and follow a *one-vs-all* strategy [27] to discriminate between multiple classes. To estimate the classification confidences, we learn a probabilistic model based on the feature vectors distance to the hyperplane for each binary SVM [28]. Finally, we combine the binary confidence estimates using a pairwise coupling scheme [29] in order to calculate the posterior probability density over all classes which forms the input for the intention recognition module.

D. Intention Recognition

Intention recognition integrates the activity and motion recognition as well as domain knowledge, e.g., time or object presence. The recognition of intentions, as e.g., the aim to *Lay Table*, is phrased as a problem of modeling, learning, and inference in *Hybrid Dynamic Bayesian Networks* (HDBN) [30], [1] as these allow for causal modeling, consistent uncertainty processing, the use of continuous- and discrete-valued variables and nonlinear dependencies [1]. The use of HDBN facilitates an extension of our approach, as inference for any HDBN may be performed generically–as long as the subsystems provide a posterior probability distribution.

Representation: The human's rationale is modeled in a discrete time HDBN. For continuous- and discrete-valued random variables, the probability densities are uniformly represented as continuous density functions $f(\underline{x})$, cf. [1]. The causal model used by the intention recognition is shown in Fig. 3. The intentions $\underline{\boldsymbol{i}}_t$ drive the human's behavior, which is modeled two-fold: as coarse activities $\underline{\boldsymbol{a}}_t$ and as fine-grained motions $\underline{\boldsymbol{m}}_t$. The activities and motions are based on distinct features $\underline{\boldsymbol{v}}_t$ and $\underline{\boldsymbol{v}}_t^{'}$ as these are post-processed differently. The parts of the model in Fig. 3 corresponding to the activity and motion recognition will not be formed explicitly but substituted by the respective subsystems' measurement updates. Extending the model in Fig. 3 with more recognizers may be easily performed by just appending random variables to the HDBN. For example, domain knowledge was introduced in the experiments by appending a binary random variable for each object class in the scenario. The binary values encode the presence or absence of objects of the respective class. Inference in the extended model is performed by standard methods described in the following.

Recognition: Inference in the HDBN of Fig. 3 requires the processing of asynchronous batch measurements from the different smoothing methods used in the activity and motion recognition. Representative for all components we consider only the motions $\underline{\boldsymbol{m}}_t$. We assume measurements $\hat{\underline{v}}_{a:0}$, $a < t$ to be given. When a new estimate $f(\hat{\underline{v}}_{b:a}|\underline{m}_b)$ for a batch of measurements $\hat{\underline{v}}_{b:a}$, $a < b < t$ is produced by the subsystem, the intention estimate is calculated as

$$f(\underline{i}_t|\hat{\underline{v}}_{b:0}) \approx \int_{\Omega_{t:a}} \int_{\mathcal{M}_b} c \cdot \overbrace{f(\underline{i}_{t:b+1}|\underline{i}_b)}^{\text{prediction}} \cdot \underbrace{[f(\hat{\underline{v}}_{b:a}|\underline{m}_b) f(\underline{m}_b|\underline{i}_{b:a})]}_{\text{measurement update}} \underbrace{\cdot f(\underline{i}_a|\hat{\underline{v}}_{a:0})}_{\text{previous filtering}} \; \mathrm{d}\underline{m}_b \, \mathrm{d}\underline{i}_{t:a} \,, \quad (1)$$

with $f(\underline{i}_{t:b+1}|\underline{i}_b) = \prod_{l=b+1}^{t} f(\underline{i}_l|\underline{i}_{l-1})$, $f(\underline{m}_b|\underline{i}_{b:a}) = f(\underline{m}_b|\underline{i}_b) \prod_{l=a+1}^{b} f(\underline{i}_l|\underline{i}_{l-1})$ and $c = f(\hat{\underline{v}}_{b:a}|\hat{\underline{v}}_{a:0})$. The estimate is approximate, as temporal dependencies between the subsystems in the HDBN, i.e., the relations between the activity and motion estimates, are neglected. If no measurements are made, quasi-stationarity is assumed, i.e. only prediction is used.

Learning: The parameters of the measurement systems, i.e., $f(\underline{a}_t|\underline{i}_t)$ and $f(\underline{m}_t|\underline{i}_t)$, are learned from training data. Because $\underline{i}_t$, $\underline{a}_t$, and $\underline{m}_t$ are discrete-valued variables, the labeled video sequences were used as completely observable data to obtain the maximum log-likelihood estimates for these conditional density functions from the sample statistics. These statistics are averages of the probability distributions over the different activities and motions as produced by the activity and motion recognition over all video frames for a given intention over all persons. The remaining conditional density functions were obtained from expert knowledge. In order to smooth the estimate sequence $\underline{\boldsymbol{i}}_{t:0}$, $f(\underline{i}_t|\underline{i}_{t-1})$ corresponds to a damping matrix. Thus $\underline{\boldsymbol{i}}_t$ converges toward a uniform distribution in the absence of measurements.

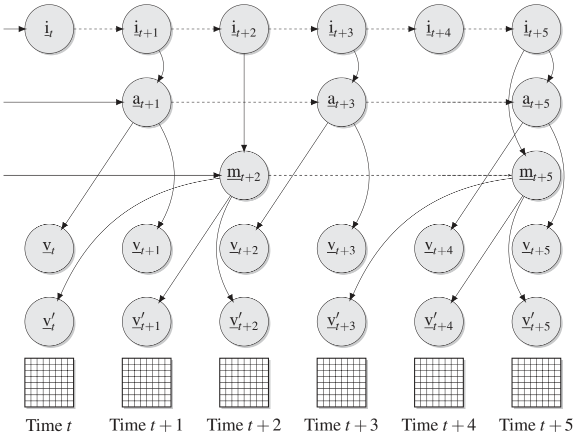


Fig. 3. An HDBN modeling the hidden intentions $\underline{\boldsymbol{i}}_t$, motions $\underline{\boldsymbol{m}}_t$, and activities $\underline{\boldsymbol{a}}_t$ as well as the observed image features $\boldsymbol{v}_t^{(,)}$ for each time step t. The dependencies of $\underline{\boldsymbol{v}}_{t+5:t}$ and $\underline{\boldsymbol{v}}'_{t+5:t}$ show the asynchronous measurements from the smoothing sub-systems. For simplicity, the domain knowledge was omitted, but may be trivially added to the HDBN.

IV. Experiments

Our experiments demonstrate the recognizers' performance at motion, activity, and intention level, the quality of the complementary recognition results and therefore, the robustness of the overall system against singular classifier failure. To show the full capabilities of our systems, we needed a data set challenging to each level of recognition. This corresponds to a lot of variation in all levels of recognition, e.g., we need various motions and activities as well as varying times of day and objects, to estimate the human's intentions. To the best of our knowledge there is no such data set. The data set[2] we collected is described in the following.

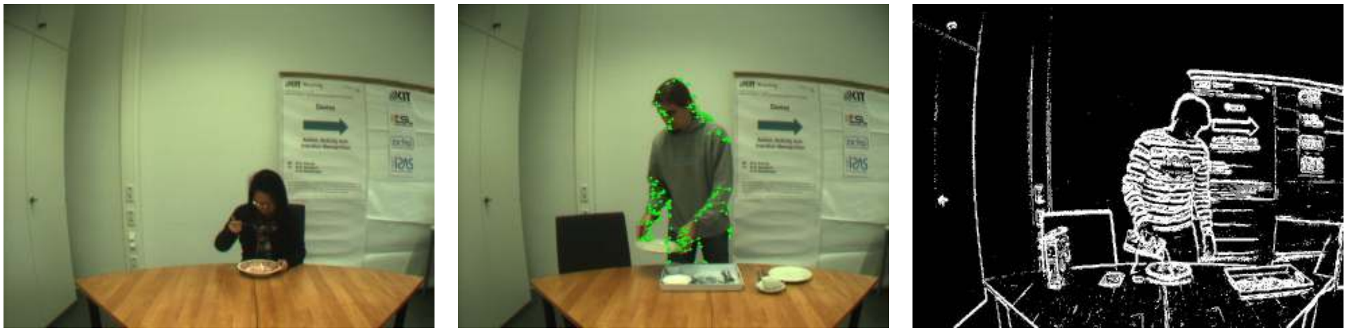
A. Hardware Setup

For the acquisition of our data set, we used a single video camera in a setup that resembles the application of our system on a humanoid robot. We imagine the robot to act as a "butler" observing the scene from a position that does not obstruct the human and offering his service whenever he assesses that it might be appreciated. For this reason, the camera view-point was fixed during the recordings to a place in front of a kitchen table, i.e., opposite to the human, cf. Fig. 4. A Point Grey Dragon-Fly Camera with a resolution of 640×480 pixels and a frame rate of 30 fps was used. In the experiments, a mix of artificial and day light (9 AM to 8 PM) as well as textured and plain background was used.

B. Scenario

The data set was collected in a kitchen setting, where ten different persons performed seven kitchen tasks. For each task, the person entered the scene, performed manipulations at the table, and left the scene afterwards, as shown in Fig. 4. The seven recorded tasks were: *Lay Table*, *Prepare Cereals*,

[2] Available at http://www.sfb588.uni-karlsruhe.de/minta/



(a) Eat with Spoon

(b) Lay Table

(c) Prepare Cereals

Fig. 4. Snapshots of three exemplary image sequences out of the set of seven used tasks *Prepare Cereals*, *Prepare Pudding*, *Lay Table*, *Eat with Spoon*, *Eat with Fork*, *Clear Table*, and *Wipe Table*. The snapshots show (a) a raw image (b) a raw image overlaid with sparse optical flow and (c) an image solely with gradient features.

Prepare Pudding, *Eat with Spoon*, *Eat with Fork*, *Clear Table*, and *Wipe Table*. The intentions combine the activities with motions and domain knowledge. In this scenario, nine intentions were used: *Lay Table*, *Prepare Cereals*, *Prepare Pudding*, *Spoon Breakfast*, *Spoon Lunch*, *Cut Breakfast*, *Cut Lunch*, *Clear Table*, *Wipe Table*, which differentiate the tasks by object and time knowledge. We denote activities as tasks, that can be discerned without the need of explicit object knowledge. Hence, the tasks *Prepare Cereals* and *Prepare Pudding* are considered one activity, i.e., *Prepare Meal*, resulting in a total of six activity classes. For a fine-grained recognition of the performed tasks a set of 60 motion primitives, e.g., *Place Object on Table*, *Pour*, or *Stir*, was defined as an alphabet for the motion recognition system. The data set was then manually annotated with the motion primitives for training and as ground truth for the recognition experiments. Although the motion recognition system does not recognize objects directly, the used objects can be recognized implicitly through the performed motion and its context. Every person performed each task ten times resulting in a total of 700 image sequences.

C. Assessment Criteria

For the evaluation of our systems, we optimized all recognizers on 560 image sequences of eight persons using 8-fold leave-one-out cross validation (LOO-CV). The 140 sequences of the two remaining persons were used as an evaluation set. Recognition results are given as the average recognition rates for the cross validation and the recognition rate on the evaluation set (EVAL set). For our experiments, the motion and activity recognition systems have been trained and optimized to give good recognition results on the motion and activity level. These results are assumed to be close to the optimal input for the intention recognition.

D. Validation of the Motion Recognition

The recognition rate of the motion recognition system is measured in terms of motion primitive accuracy (ACC):

$$\mathrm{ACC} = (1 - \frac{\#\mathrm{ins} + \#\mathrm{del} + \#\mathrm{sub}}{\#\mathrm{primitives\ in\ reference}}) \times 100\% \,. \qquad (2)$$

We compared the recognizer output (sequence of the most likely motion primitives, which are passed to the intention recognition) with the manually annotated sequences. Motion recognition results of the LOO-CV and on the EVAL set for off-line and on-line recognition are reported in Tab. I. For the off-line recognition the recognition process uses all images of a motion sequence at a time, which allows a better normalization of the features. For the on-line recognition, the images are processed directly and are never considered again. For the evaluation of the intention recognition, we used the *on-line* results. Due to the worse feature normalization, the on-line recognition results with an accuracy of 58.6 % are worse than the off-line results, but the system outputs the recognized motions with a lot shorter response time. The recognition rates on the 60 motion primitives for the 2 configurations are given in Tab. I. Note that the chance of randomly guessing the result correctly is $\frac{1}{60}$ here. Fig. I shows the results of the motion recognition system for the different tasks. The primitive recognition rate is high for most of the tasks.

E. Validation of the Activity Recognition

We measure the performance of the activity recognition as the per-window classification rate. The parameter with the highest impact on the accuracy is the temporal size of the sliding window. On the one hand it should be as short as possible to minimize response time, but on the other it should also capture enough information to allow an accurate recognition. From experiments using the LOO-CV, we concluded that a window duration of 60 frames, corresponding to 2 seconds, yields a good trade-off between both. We regard each window to be independent from past observations with the reasoning that the robot may enter the kitchen while an activity has already started and should still be able to assess the situation.

An accumulated confusion matrix of the activity recognition results for the EVAL set is shown in Fig. 5. It can be seen that our approach is generally quite robust, but has problems to discriminate between the activities *Lay Table* and *Clear Table* with average recognition rates on the EVAL set of 44.3 % and 30.5 % (see Tab. II) respectively. This is not surprising though, as the motion patterns of both activities are very similar and thus, can be easily confused.

TABLE I

AVERAGE MOTION PRIMITIVE ACCURACY GIVEN PER TASK FOR THE LOO-CV AND EVAL SETS.

Task		Lay Table	Prepare Cereals	Prepare Pudding	Eat with Spoon	Eat with Fork	Clear Table	Wipe Table	*Avg. Rate*	*Chance*
Accuracy (off-line)	LOO-CV	76.9 %	78.9 %	76.1 %	73.1 %	58.0 %	86.5 %	47.3 %	**70.6** %	1.7 %
	EVAL	83.1 %	87.5 %	80.4 %	42.5 %	72.8 %	89.7 %	67.0 %	**74.4** %	1.7 %
Accuracy (on-line)	LOO-CV	62.4 %	61.2 %	59.5 %	66.1 %	44.8 %	59.8 %	42.4 %	**56.3** %	1.7 %
	EVAL	66.3 %	65.3 %	57.0 %	44.7 %	63.1 %	61.3 %	55.5 %	**58.6** %	1.7 %

F. Validation of the Intention Recognition

The performance measure for the intention recognition is the consistency of the ML estimate with the ground truth. The performance was evaluated for the recorded image sequences for every intention, where the intention was estimated every 2nd frame and then compared with the ground truth. In order to test the robustness against missing and delayed measurements, the results of the activity (motion) recognition were integrated every 4th (30th) frame, respectively. These rates are arbitrary and may be increased to an integration of all measurements in each frame. A uniform prior distribution was used and the domain knowledge was set according to the ground truth. Therefore, the intention estimate is a uniform distribution until the first measurement arrives. The uniform distribution is considered as a misclassification. The uncertainty of the object knowledge was set to a 75% combination of perfect information and a uniform distribution. Fig. 7 gives the recognition rates for the intention recognition w.r.t. the EVAL set in terms of correct ML estimate with varying component setup, i.e., only the domain knowledge and the activity recognition, only the domain knowledge and the motion recognition as well as domain knowledge, activity, and motion recognition were used. The results in Fig. 7 demonstrate that the complementarity of the activity and motion recognition improves the estimates for almost all intentions. The average classification rate for the ML intention estimate using only the domain knowledge and motion recognition is 80.3 %, using only the domain knowledge and activity recognition is 80.4 %, and using all sources of information is 83.5 %. Fig. 6 shows the probability of the ML estimate over time and how the performance improves with the advent of either recognition results and especially, when both estimates become available regularly (around frame 150). In Fig. 7, the effect of the domain knowledge can be seen in the high recognition rates for the different breakfast and lunch types. Without uncertain information about the time of day and object presence, these would not be distinguishable by mere activity and motion.

G. Results

The results presented in Sec. IV-D-IV-F for each recognition level, the results in Fig. 7, and the recognition probabilities over time, as shown in Fig. 6, clearly demonstrate the quality and advantage of the multi-level approach. Especially Fig. 6 and 7 visualize the different measurement frequencies and complementary contributions of both mid-level recognition results. The fusion of the mid-level results in the intention recognition not only increases the recognition rate, but allows for further distinction of intentions by adding object and time knowledge. Due to the modularity of the approach, it is hard to give exact run-times for the entire system. Each of the recognition systems consumes less than 30 ms per frame, rendering an on-line and real-time application of the system tractable even for much larger scenarios.

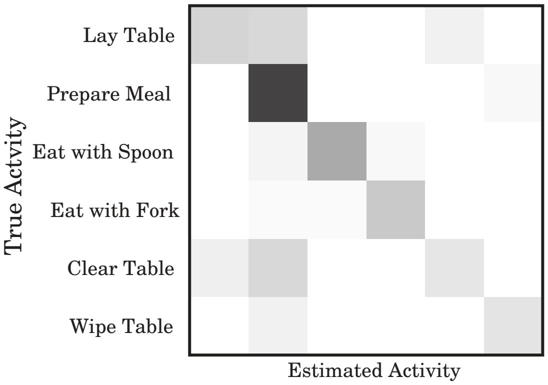


Fig. 5. Accumulated confusion matrix of the activity recognition for the EVAL set, corresponding to an average recognition rate of 67.2%. The gray values correspond to normalized frequencies.

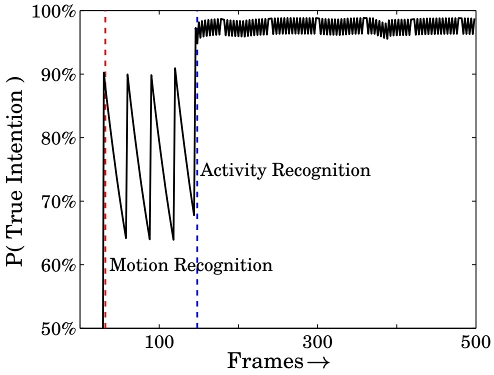


Fig. 6. Probability of the intention estimate for *Spoon Breakfast* over the first 500 frames. The frame of each first recognized activity and motion is marked with a dashed line.

V. CONCLUSIONS AND FUTURE WORKS

A multi-level approach to intention, activity, and motion recognition was proposed. Based on monocular video input, the recognition is performed on-line and in real-time. The system is limited to fixed view-points, but is independent of the test person. The main contributions are the extensible, multi-level modeling, the efficient activity and motion recognition, and the information fusion based

TABLE II

AVERAGE RECOGNITION RATES GIVEN PER ACTIVITY FOR THE LOO-CV AND EVAL SETS.

Activity		Lay Table	Prepare Meal	Eat with Spoon	Eat with Fork	Clear Table	Wipe Table	*Avg. Rate*	*Chance*
Recognition	LOO-CV	46.9 %	86.8 %	81.6 %	80.6 %	48.3 %	73.9 %	**69.7** %	16.7 %
Rate	EVAL	44.3 %	94.5 %	82.1 %	88.5 %	30.5 %	63.0 %	**67.2** %	16.7 %

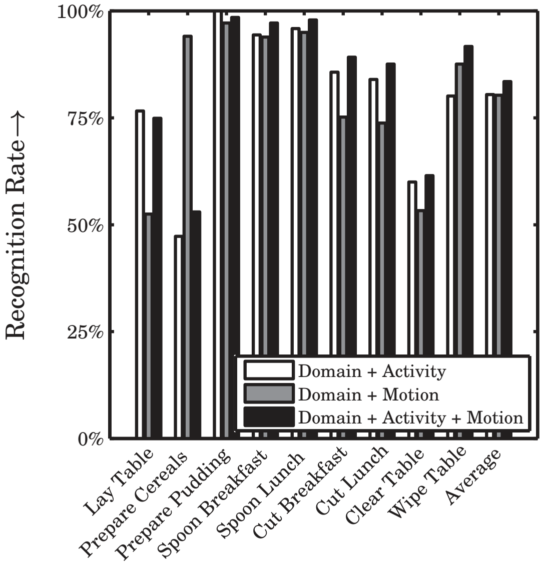


Fig. 7. Recognition rates for the intention based on the ML estimate of the intentions and differing components setup: given domain knowledge and activity recognition only, given domain knowledge and motion recognition only, and given domain knowledge, activity as well as motion recognition.

on generic processing of asynchronous recognition results. We performed experiments on a corpus of complex kitchen tasks containing a mix of artificial and day light as well as textured and plain background. The results are promising and show the robustness of the entire recognition system against singular classifier failure. As future work, a larger set of view-points and the incorporation of a vision-based object recognition is considered in order to obtain a fully integrated and stand-alone system. The performance of long-term and non-stop usage needs to be evaluated. The system will be integrated with a larger multi-modal dialog system and will become part of the humanoid robot ARMAR [31].

REFERENCES

[1] O. C. Schrempf, A. Hanselmann, and U. D. Hanebeck, "Efficient Representation and Fusion of Hybrid Joint Densities for Clusters in Nonlinear Hybrid Bayesian Networks," in *Fusion*, 2006.

[2] B. Schölkopf and A. Smola, *Learning with Kernels: support vector machines, regularization, optimization, and beyond*, 2002.

[3] L. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, 1989.

[4] O. C. Schrempf and U. D. Hanebeck, "A Generic Model for Estimating User Intentions in Human-Robot Cooperation," in *ICINCO*, 2005.

[5] S. Carberry, "Techniques for Plan Recognition," *User Modeling and User-Adapted Interaction*, 2001.

[6] H. H. Bui, "A General Model for Online Probabilistic Plan Recognition," in *IJCAI*, 2003.

[7] K. A. Tahboub, "Intelligent Human-Machine Interaction Based on Dynamic Bayesian Networks Probabilistic Intention Recognition," *Intelligent and Robotic Systems*, 2006.

[8] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine Recognition of Human Activities: A Survey," *IEEE Transactions on Circuits and Systems*, 2008.

[9] R. Poppe, "A Survey on Vision-Based Human Action Recognition," *Image and Vision Computing*, 2010.

[10] S. Park and J. K. Aggarwal, "A Hierarchical Bayesian Network for Event Recognition of Human Actions and Interactions," *Multimedia Systems*, 2004.

[11] Y. A. Ivanov and A. F. Bobick, "Recognition of Visual Activities and Interactions by Stochastic Parsing," *PAMI*, 2000.

[12] S. D. Tran and L. S. Davis, "Event modeling and recognition using markov logic networks," in *ECCV*, 2008.

[13] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid, "Evaluation of Local Spatio-Temporal Features for Action Recognition," in *BMVC*, 2009.

[14] J. Yang, Y. Xu, and C. S. Chen, "Human Action Learning via Hidden Markov Model," *IEEE Trans. on Systems, Man, and Cybernetics*, 2002.

[15] M. Brand, N. Oliver, and A. Pentland, "Coupled Hidden Markov Models for Complex Action Recognition," in *CVPR*, 1997.

[16] D. Gehrig, T. Stein, A. Fischer, H. Schwameder, and T. Schultz, "Towards semantic segmentation of human motion sequences," in *Proceedings of the 33rd Annual German Conference on Advances in Artificial Intelligence*, 2010.

[17] A. Casile and M. A. Giese, "Critical Features for the Recognition of Biological Motion," *Journal of Vision*, 2005.

[18] H. Soltau, F. Metze, C. Fügen, and A. Waibel, "A One-Pass Decoder based on Polymorphic Linguistic Context Assignment," *ASRU*, 2001.

[19] M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, and M. Westphal, "The Karlsruhe-Verbmobil Sspeech Recognition Engine," *ICASSP*, 1997.

[20] C. Nevill-Manning and I. Witten, "Identifying Hierarchical Structure in Sequences: A linear-time algorithm," *Journal of Artificial Intelligence Research*, 1997.

[21] T. M. Martinetz, S. G. Berkovich, and K. J. Schulten, "'Neural-Gas' Network for Vector Quantization and its Application to Time-Series Prediction." *IEEE Transactions on Neural Networks*, 1993.

[22] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual Categorization with Bags of Keypoints," in *ECCV Workshop on Statistical Learning in Computer Vision*, 2004.

[23] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning Realistic Human Actions from Movies," in *CVPR*, 2008.

[24] K. Schindler and L. Van Gool, "Action Snippets: How many frames does human action recognition require?" in *CVPR*, 2008.

[25] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up Robust Features," in *ECCV*, 2006.

[26] R. Poppe and M. Poel, "Comparison of Silhouette Shape Descriptors for Example-based Human Pose Recovery," in *Intl. Conf. on Automatic Face and Gesture Recognition*, 2006.

[27] C.-C. Chang and C.-J. Lin, "LIBSVM: a Library for Support Vector Machines," 2001.

[28] H.-T. Lin, C.-J. Lin, and R. C. Weng, "A note on Platt's Probabilistic Outputs for Support Vector Machines," *Machine Learning*, 2007.

[29] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability Estimates for Multi-class Classification by Pairwise Coupling," *Journal of Machine Learning Research*, 2004.

[30] K. Murphy, "Dynamic Bayesian Network: Representation, Inference and Learning," Ph.D. dissertation, UC Berkeley, 2002.

[31] T. Asfour, K. Regenstein, P. Azad, J. Schröder, A. Bierbaum, N. Vahrenkamp, and R. Dillmann, "ARMAR-III: An integrated humanoid platform for sensory-motor control," in *Humanoids*, 2006.