Towards Automated Models of Activities of Daily Life

Michael Beetz, Jan Bandouch, Dominik Jain and Moritz Tenorth
Intelligent Autonomous Systems, Technische Universität München
{beetz, bandouch, jain, tenorth}@cs.tum.edu

***Abstract*— We propose *automated probabilistic models of everyday activities (AM-EvA)* as a novel technical means for the perception, interpretation, and analysis of everyday manipulation tasks and activities of daily life. AM-EvAs are based on action-related concepts in everyday activities such as action-related places (the place where cups are taken from the cupboard), capabilities (the objects that can be picked up single-handedly), etc. These concepts are probabilistically derived from a set of previous activities that are fully and automatically observed by computer vision and additional sensor systems. AM-EvA models enable robots and technical systems to analyze activities in the complete situation and activity context. They render the classification and the assessment of actions and situations objective and can justify the probabilistic interpretation with respect to the activities the concepts have been learned from. In this paper, we describe the current state of implementation of the system that realizes this idea of automated models of everyday activities and show example results from the observation and analysis of table setting episodes.**

I. Introduction

Enabling ambient environments and autonomous robots to competently interpret and analyze everyday activities requires the systematic and comprehensive observation of the activities and the abstraction of the observed behavior into informative models (Fig. 1). Such models are to enable robots to infer the intentions of people, the abnormality of the behavior, the next actions, where people go, why an action failed, the most likely trajectory of a reaching motion, etc.

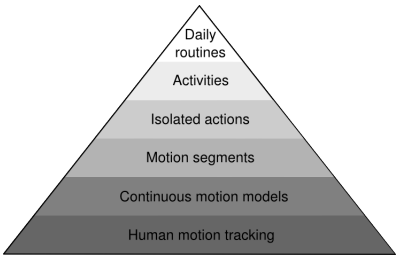

Fig. 1. Abstraction from raw sensor data to description of a daily schedule.

We propose the use of probabilistic models of everyday activities in order to gracefully deal with the uncertainty and variation inherent in human activities. However, the effective realization of these probabilistic models poses several hard research challenges:
(1) *Expressiveness*: Probabilistic models of everyday manipulation must represent continuous motion as well as discrete events. They also need to be *relational* in order to represent behavior that is determined by the circumstances and interactions with individual objects. Learning such expressive models and performing inference effectively as well as efficiently is well beyond the current state of the art.
(2) *Groundedness in sensor observations*: Concepts representing aspects of everyday activity should be specified with respect to the roles they play in people's activities. For example, the system could learn the places in the environment where people stand when taking objects out of the cupboard.
(3) *Automation through observation systems and action interpreters*: Activity models should be automatically constructed from highly reliable and accurate observation systems for long activity episodes. In addition, high-dimensional time series data (e.g. 51 DOF human pose sequences) has to be abstracted into compact and informative representations.

In the AM-EvA project, we investigate a new generation of activity models that

- are based on full human body motion and object interactions as their primitive building blocks;
- represent the interaction between navigation and manipulation actions, the objects they are operating on, the situation context and the location, and thereby allow for more comprehensive assessment of the activities;
- use concepts such as the place where actions are performed, the objects that can be picked up single-handedly, etc. These concepts are defined transparently and therefore constitute objective criteria for classification and assessment;
- can be acquired automatically by a camera- and sensor-network-based observation system.

The main objectives of the AM-EvA project are the investigation of novel computational mechanisms that enable computer systems to recognize intentional activities, the development of an integrated software system to automate activity analysis, and the demonstration of the impact of automated activity analysis on service robotics and ambient assistive living environments.

II. Overview

The software architecture of the AM-EvA system is depicted in Fig. 2. Observation data is provided by a distributed camera and sensor network that includes a highly accurate markerless full-body motion tracker and perceives interactions with objects, such as pick up and put down, based on RFID sensors. The system interprets the time series data by segmenting it into semantically meaningful motion segments. These include continuous data such as reaching trajectories, and discrete events such as *making contact with an object*.

Actions are represented in a first-order logical language based on time intervals where relations on continuous data are processed lazily. This means predicates operating on continuous motion data such as the velocity profile of a

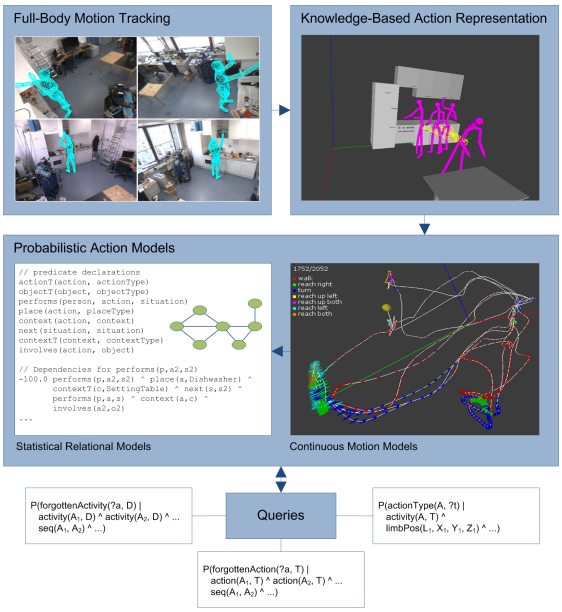

Fig. 2. Software architecture of the AM-EvA system.

reaching motion are computed on demand and then translated into their symbolic representation.

Sequences of motion segments are combined into intentional actions, such as picking up an object or opening a door. The activity model is then a joint probability distribution over this first-order action representation language, which we represent as Markov logic and Bayesian logic networks. It can be queried by users and other system components.

We model activities of daily life at different levels of abstraction: A *routine level* that includes a complete daily schedule, an *activity level*, which models complete activities (e.g. setting a table, preparing a meal), and an *action level* that models actions (e.g. picking up an object) as a probabilistic hybrid discrete/continuous event (Fig. 1). Lower levels in the pyramid are a sequence of *motion segments*, *continuous motion models* and the raw output of the *human pose tracking system*. This strategy allows to detect many different kinds of abnormal behavior without modeling each and every action in excessive detail.

The (semi-)automatically acquired model enables the system to automatically infer answers to a large variety of queries concerning the performance of everyday activities including the following ones: Did the human go for his daily walk? Did he take his medicine at the correct time? Is he still able to cook meals or is he forgetting things? Does the human have problems reaching into the overhead cupboards?

III. Observing Everyday Manipulation Activities

We believe that observing human activities requires the estimation of human motions and associated full-body poses. However, commercial marker-based tracking systems are infeasible in real scenarios, as they are intrusive, expensive and difficult to set up. We developed a markerless motion capture system suitable for everyday environments (Fig. 3) that comes with the following improvements over compareable state-of-the-art systems: (1) Setup is easy, cheap and unintrusive, requiring only the placement of 3 or more cameras; (2) Estimation of 51 DOF poses (joint angles) and corresponding body part trajectories with high accuracy; (3) Unconstrained tracking without prior training; (4) Implicit modeling of objects and the environment.

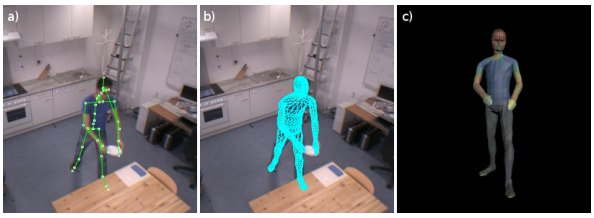

Fig. 3. Markerless motion capture (one of four cameras): a) inner model b) outer model c) virtual 3D view with appearance model. This and other videos are available at `http://memoman.cs.tum.edu`.

Our system estimates (previously unobserved) human poses in a recursive Bayesian framework using a variant of particle filtering. To make the problem computationally tractable, we developed a sampling strategy [2] that is a combination of search space partitioning [3] with a multi-layered search strategy [4]. By integrating the sampling strategy with a silhouette-based multi-camera tracking framework and an anthropometric human model [1], we are able to derive realistic human posture at low frame rates (25 Hz). Evaluations on the *HumanEvaII* benchmark [5] show the accuracy and robustness of our approach (Fig. 4).

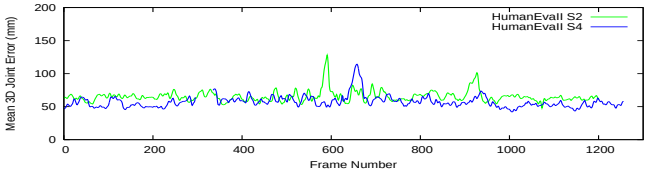

Fig. 4. Motion capture accuracy on *HumanEvaII* benchmark [5].

To be able to track subjects performing manipulation tasks such as picking up objects from inside a cupboard and placing them on a table (Fig. 3), dynamic parts in the environment are filtered based on learnt human appearance, and ignored when evaluating particle weights. Occlusions from static environment objects (e.g. tables) are handled by using blocking layers that prevent evaluation unless the blocked area resembles human appearance. Good tracking results are achieved when every occluded part is visible by at least 2-3 other cameras.

IV. Continuous Motion Models

At the most detailed level, our models base their representations directly on the joint motions that are gathered by the markerless full-body motion tracking system (together with information on object interactions from the sensor network in general). Even though the data at this level is very high-dimensional, it is reasonable to assume that it is nevertheless well-structured, because actions performed during housework are in many ways constrained (with respect to the expected limb motions, which are far from arbitrary) and they follow clearly discernable patterns. These patterns can be made explicit by suitably embedding the high-dimensional data into a low-dimensional latent space.

We can directly incorporate the semantic labellings of action sequences as an input dimension to the learning algorithms of, for example, Gaussian process dynamical models (GPDMs), and learn low-dimensional embeddings

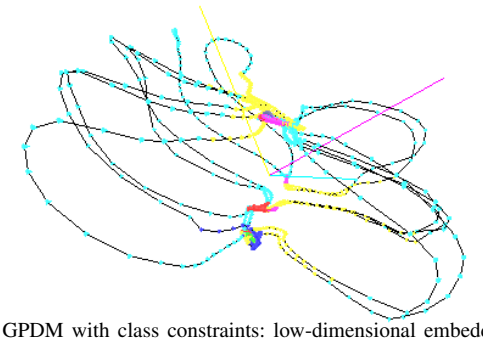

Fig. 5. GPDM with class constraints: low-dimensional embedding of an unconstrained pick-and-place task; the colors correspond to semantic labels.

that seek to structure the latent space with respect to the labels. We thus extended learning algorithms for GPDMs with probabilistic constraints that ensure that points belonging to the same classes are close together while points belonging to different classes may be far apart, further structuring the latent space according to its semantic interpretation (see Fig. 5). (A somewhat similar approach, which, focusing on specific inference tasks, considers a discriminative model, was proposed in [7].) Given a learnt mapping from the high-dimensional data space to the latent space, we can then perform classification of newly observed sequence data by maximizing the likelihood of the latent space configuration given the labels.

Since our models are generative, we can flexibly use them to either predict future motions that are likely to occur, evaluate the probability of an observed motion sequence (allowing us to detect peculiar actions/motions that are, for example, unusual given the overall actions that are supposed to be performed) or, as previously stated, infer the labels of a sequence, providing the discretized information for higher-level modeling. The labelling constitutes precisely the semantic interpretation that we require to analyze activities at higher levels of abstraction.

In addition to the aforementioned approaches based on continuous models of the immediate motions, we also use linear-chain conditional random fields (CRFs) in order to identify motion segments and represent them as instances of classes such as *Reaching* or *TakingSomething*. The approach is somewhat complementary, as it does not result in motion models, yet it facilitates the incorporation of further sensory input such as RFID readings, allowing semantic labellings to be more easily based on object interactions.

As input, we use a quite large set of binary features, which could be split into two groups: The first group are pose-related features which denote, for example, if the human is extending or retracting the hands. The second group of features includes information from the environment model and the sensor network and states for instance if the human currently carries an object or if a drawer is being opened. These features are combined, and CRFs are learned on a labeled training set. The CRF classifiers are integrated into the knowledge processing system and create a sequence of motion segments which are represented as instances of the respective semantic classes. The system is able to classify unknown test sequences with an accuracy of up to 89% in spite of the significant variation in how the actions are performed.

V. Symbolic Action Representation

On the next abstraction layer, the result of the segmentation is represented as a sequence of motion segments of the respective classes as depicted in the lower part of Fig. 6. The symbolic action representation is part of a larger knowledge representation system [8]. Motion segments are represented as instances of motion classes like *Reaching*, inheriting all their properties. The pyramid of higher-level actions in Fig. 6 is built by matching the sequence of motion segments to the descriptions of actions and their sub-events. Action parameters are determined by relating the segments to events observed simultaneously from the sensor network, like an object being picked up or a cupboard being opened. The observations are automatically loaded into the knowledge processing system using *computable* classes and properties which are also described in [8]. Examples of queries can be found in Fig. 7. Though the traces are only drawn for a single joint, the result includes the full human pose vectors for each point in time. In the image on the left, we asked for the whole pose sequence of a table setting activity using the following query:

```
owl_query(?Acty, type, 'SetTable'),
postureForAction(?Acty, ?Posture)
```

This query directly relates the activity level and the pose vectors, but a more fine-grained selection that takes the intermediate levels of abstraction and different action parameters into account is possible as well. For instance, the query depicted in the right image in Fig.7 asks for all postures that are part of a *TakingSomething* motion, performed on a *DinnerPlate* in a *TableSetting* context:

```
owl_query(?Acty, type,          'SetTable'),
owl_query(?Actn, type,          'TakingSomething'),
owl_query(?Actn, subEvent,       ?Acty),
owl_query(?Actn, objectActedOn, ?Obj),
owl_query(?Obj,  type,          'DinnerPlate'),
postureForAction(?Actn, ?Posture)
```

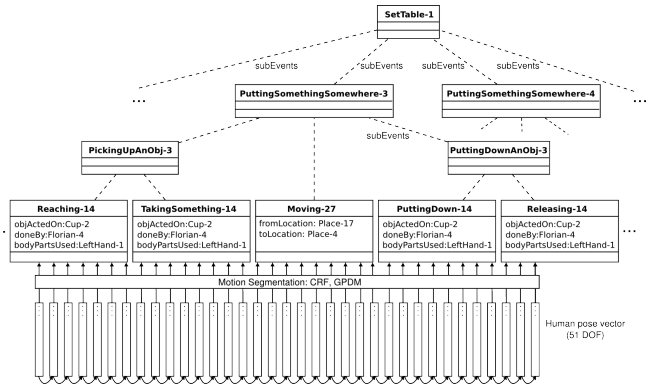

Fig. 6. Hierarchical action model constructed from observed human tracking data.

As described in [8], the data can also be used for learning the place from which an action is usually performed. Fig. 8 sums up the main steps: The observed positions are loaded into the knowledge processing system (left) and clustered with respect to their Euclidean distance (center). These clusters are represented as "places" in the knowledge base, and the system automatically learns a mapping from action properties (like the object to be manipulated and its position)

to the place where the human is standing. Using this model, it is possible to either obtain a place for an action (like the place for picking up pieces of tableware drawn in Fig. 8 right) or to classify observations in order to find out about the most probable action performed from this place.

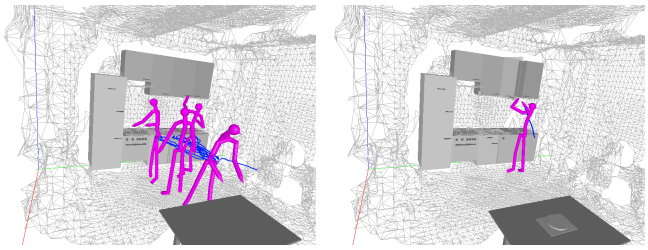

Fig. 7. Human pose sequences for setting a table (left) and taking a plate out of the cupboard (right).

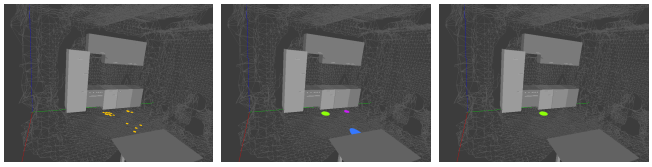

Fig. 8. Observed positions of manipulation actions (left), the positions clustered into places (center), and the result that has been learned as the "place for picking up pieces of tableware" (right).

VI. Activity Models

We consider sequences of single (*atomic*) actions, such as setting a table or cooking meals, as activities. Spotting differences at this level requires a detailed modeling of the actions involved, their parameters, a (potentially partial) ordering between them, etc.

Hierarchical action models (Fig. 6) that combine logical activity models with probabilistic logical reasoning provide a detailed representation of single actions, covering the whole range from the activity level (cooking pasta), over single actions (switching on the stove, putting salt into the water) to motion segments (the arm movement for taking the pasta box out of the cupboard). Apart from just the pure action sequences, the models provide information about manipulated objects, hands used, places where people stand during actions, or the purpose of the action.

To recognize activities, the system needs detailed descriptions of involved actions and their parameters (e.g. temporal extent, objects being manipulated, spatial location). Our system autonomously expands its repertoire of actitivies by importing information from the web. We integrated methods for importing activity specifications from sites like ehow.com and for transforming these natural-language task instructions into formal logic descriptions [9]. Using this approach, we have successfully classified test sequences as instances of different activities using description logic reasoning.

VII. Models of the Daily Schedule

The most abstract level describes whole days as sequences of activities, developing models of the usual daily schedule. The models have to describe and allow the detection of common activities like getting up, cooking, having meals, going for a walk, reading, watching TV, or taking a bath.

Probabilistic logical representations provide both the required expressiveness and flexibility. Technically, the models are very similar to those on the activity level, the only major difference being that they are applied to more abstract entities. Accordingly, peculiarities that can be detected with such models are on the level of complete activities, for instance that a person omitted activities that would usually take place, such as leaving the apartment.

VIII. Conclusions

We have described the current state of AM-EvA, *automated probabilistic models of everyday activities* for the perception, interpretation, and analysis of everyday manipulation tasks and activities of daily life. We have outlined the main components, which are a full-body motion tracking system for people performing everyday manipulation tasks, various learning approaches that infer low-dimensional activity representations from the tracking data and segment continuous motions into hybrid automata representations. Other components combine these hybrid activity models with encyclopedic knowledge about everyday manipulation tasks and human living environments to provide prior knowledge for making learning and inference more tractable. AM-EvA seemlessly combines symbolic knowledge with observed behavior data structures through the use of computable relations and properties that are evaluated directly on AM-EvA's data structures and through data mining mechanisms that learn symbolic concepts (such as objects that can be picked up single-handedly) from observed activity data.

In our ongoing research, we investigate various challenges that are raised by our approach, i.e. full-body motion tracking with integrated object interaction estimation, accurate real-time tracking for high DOF motion using dimension reduction techniques, effective probabilistic learning and reasoning mechanisms for relational hybrid (discrete-continuous) action models, and high-level representations for the interleaved execution of everyday activities.

IX. Acknowledgments

This work is supported in part within the DFG excellence initiative research cluster *Cognition for Technical Systems – CoTeSys*, see also `www.cotesys.org`.

References

[1] J. Bandouch, F. Engstler, and M. Beetz, "Accurate human motion capture using an ergonomics-based anthropometric human model," in *Proc. of the 5th International Conference on Articulated Motion and Deformable Objects (AMDO)*, 2008.

[2] J. Bandouch, F. Engstler, and M. Beetz, "Evaluation of hierarchical sampling strategies in 3d human pose estimation," in *Proc. of the 19th British Machine Vision Conference (BMVC)*, 2008.

[3] J. MacCormick and M. Isard, "Partitioned sampling, articulated objects, and interface-quality hand tracking," in *Proc. of the 6th European Conference on Computer Vision (ECCV)*, 2000.

[4] J. Deutscher and I. Reid, "Articulated body motion capture by stochastic search," *International Journal of Computer Vision (IJCV)*, vol. 61, no. 2, pp. 185–205, 2005.

[5] L. Sigal and M. J. Black, "Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion," Brown University, Tech. Rep., 2006.

[6] J. Wang, D. Fleet, and A. Hertzmann, "Gaussian process dynamical models," *Advances in Neural Information Processing Systems*, vol. 18, p. 1441, 2006.

[7] R. Urtasun and T. Darrell, "Discriminative Gaussian process latent variable model for classification," in *Proc. of the 24th International Conference on Machine Learning (ICML)*, 2007.

[8] M. Tenorth and M. Beetz, "KnowRob — Knowledge Processing for Autonomous Personal Robots," 2009, iEEE/RSJ International Conference on Intelligent RObots and Systems. To appear.
[9] M. Tenorth, D. Nyga, and M. Beetz, "Understanding and executing instructions for everyday manipulation tasks from the world wide web," IAS group, Technische Universität München, Fakultät für Informatik, Tech. Rep., 2009.