The TUM Kitchen Data Set of Everyday Manipulation Activities for Motion Tracking and Action Recognition

Moritz Tenorth, Jan Bandouch, Michael Beetz
Intelligent Autonomous Systems Group, Department of Informatics
Technische Universität München, Munich, Germany
{tenorth, bandouch, beetz}@cs.tum.edu
http://kitchendata.cs.tum.edu

Abstract

We introduce the publicly available TUM Kitchen Data Set as a comprehensive collection of activity sequences recorded in a kitchen environment equipped with multiple complementary sensors. The recorded data consists of observations of naturally performed manipulation tasks as encountered in everyday activities of human life. Several instances of a table-setting task were performed by different subjects, involving the manipulation of objects and the environment. We provide the original video sequences, full-body motion capture data recorded by a markerless motion tracker, RFID tag readings and magnetic sensor readings from objects and the environment, as well as corresponding action labels. In this paper, we both describe how the data was computed, in particular the motion tracker and the labeling, and give examples what it can be used for. We present first results of an automatic method for segmenting the observed motions into semantic classes, and describe how the data can be integrated in a knowledge-based framework for reasoning about the observations.

1. Introduction

More and more technologies are developed to assist humans in the household: Sensor-equipped environments for monitoring the behavior and assessing the level of independence of elderly people, or robots for performing household chores like setting a table or filling a dishwasher. For these applications, it is crucial to be able to model and recognize naturally performed human actions.

While there is a large body of work on action recognition, the recognition and understanding of *natural everyday human activities* is still difficult due to some hard challenges:

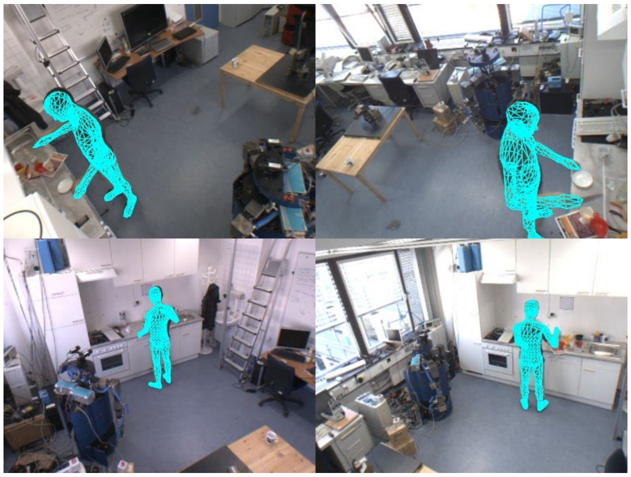

Figure 1. TUM kitchen environment. Views from the four overhead cameras with tracked human pose overlaid on top of the image

- There is much variability in how the same action can be performed, beginning from different arm postures for holding an object, different grasps to pick it up, different places and orientations where to put it, up to a varying order in which objects are manipulated.
- Object interactions can significantly change the human silhouette and impose constraints on allowed movements (e.g. when carrying a cup with hot tea).
- Humans perform several actions in parallel, using both the left and the right hand.
- Natural environments are filled with furniture, which causes occlusions of different parts of the body.
- Motions of everyday activities can be very subtle as opposed to actions such as “kicking” or “punching”.

We created the TUM Kitchen Data Set since we found that many of these challenges are not sufficiently covered in the available data sets, neither in those for motion tracking, nor in those for action recognition (a more detailed discussion of related data sets is provided in the next section). We also hope to foster research in the areas of markerless human motion capture, motion segmentation and human activity recognition. The data set will aid researchers in these fields by providing a comprehensive collection of sensory input data that can be used to develop and verify their algorithms. It is also meant to serve as a benchmark for comparative studies given manually annotated "ground truth" labels of the ongoing actions. First challenging test marks on motion segmentation have been set by us using linear-chain Conditional Random Fields (Section 5.1).

The recorded sequences consist of everyday manipulation activities in a natural kitchen environment (Fig. 1), with a focus on *realistic* motions, as opposed to artificial action sets consisting of motions such as “kick”, “punch” or “jumping jack”. We provide human motion capture data that has been tracked using our markerless tracking system for unconstrained human motions [2]. Our setup is completely unintrusive for the human subject, and the recorded video sequences are clear of obstructive technical devices (Fig. 3). The motion tracker is capable of tracking a high-dimensional human model (51 DOF) without constricting the type of motion, and without the need for training sequences. It reliably tracks humans that frequently interact with the environment, that manipulate objects, and that can be partially occluded by the environment. We have post-processed the motion capture data for the final data set whenever there was significant tracking failure. The recorded videos and the motion capture data are further complemented by RFID tag readings and magnetic sensor readings from objects and the environment.

We also present first classification results on the sequences of our data set for an automatic method for segmenting observed motions into semantic classes using linear-chain Conditional Random Fields. The classification accuracy for this challenging task reaches up to 93.1% for sequences of the same subject, and 62.77% for cross-subject classification. Furthermore, we describe how the data can be integrated in a knowledge-based framework for reasoning about the observations.

The remainder of this paper is organized as follows: After a short review of related work (Section 2), we describe the content of the TUM Kitchen Data Set and the file formats (Section 3). Section 4 then explains how the data was acquired, especially the motion tracking and labeling. In Section 5, we show examples of how the data can be used for motion segmentation and how to integrate it into a knowledge-based action interpretation framework.

2. Related Work

Research on action recognition can be divided into two directions. Holistic approaches [7, 15, 5] try to find a direct mapping between observable image cues and action labels, usually by training discriminative classifiers such as Support Vector Machines [15]. Evaluation of such approaches is often done on the KTH data set [15] or the Weizmann data set [5]. Both contain extremely articulated motions like walking, waving with the arms or “jumping jack”, that are of limited relevance to practical applications. Considering the single camera setups in these data sets, action recognition will furthermore be view-dependent.

In contrast to holistic approaches, model-based approaches to action recognition perform action classification in phase space, using parameters of a representative model that have been determined beforehand. Commonly used parameters are joint angles or joint locations of articulated human models [16, 12]. These parameters are invariant to translations, scale and rotations, making them well-suited for action recognition. However, the model parameters are difficult to extract by unintrusive means, and most of the methods rely on commercial marker-based motion capture systems for retrieving the joint parameters. Several motion-capture-only data sets such as the CMU Motion Capture Database[1] or the MPI HDM05 Motion Capture Database[2] are available that provide large collections of data. Again, the available motions are extremely articulated, well separated, and do not resemble natural everyday activities, nor do they involve manipulation or interaction tasks. The Karlsruhe Human Motion Library [1] consists of motion capture data specifically recorded for human motion imitation on humanoid robots.

A few data sets of natural motions exist that are suitable for everyday activity recognition: The CMU Kitchen Data Set[3] contains multi-modal observations of several cooking tasks, including calibrated cameras and motion capture data. Due to the large number of actions and the high variation between the actors, this data set is extremely challenging for action recognition. Furthermore, the actors are heavily equipped with technical devices, making it difficult to evaluate *e.g.* markerless motion trackers on the video data.

Uncalibrated video data and RFID readings (without motion capture data) are provided by the LACE Indoor Activity Benchmark Data Set[4], which is mainly targeted towards a coarser, general description of activities. Even coarser, higher-level data is available in the MIT House_n Data Set[5], which is aimed at recognizing activities as a whole, while we are also interested in modeling the single actions and even different motions an activity consists of.

Due to the complications associated with marker-based motion capture systems (intrusive setup, problems with occlusions, necessary post-processing, high cost) and their inapplicability in practical everyday use, computer vision research is aiming to develop markerless motion capture systems. Several systems have been presented [8, 10], yet most

[1] http://mocap.cs.cmu.edu

[2] http://www.mpi-inf.mpg.de/resources/HDM05/

[3] http://kitchen.cs.cmu.edu

[4] http://www.cs.rochester.edu/∼spark/muri/

[5] http://architecture.mit.edu/house_n/data/

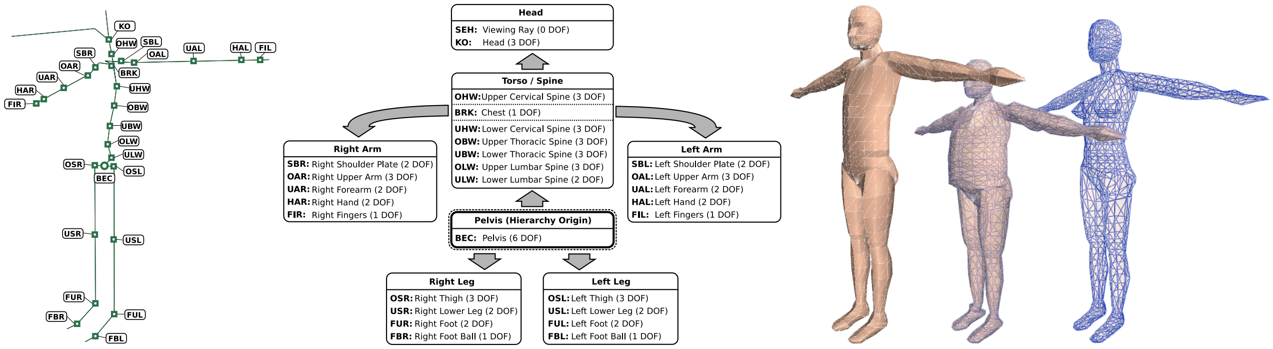
KO
OHW
SBL
OAL
UAL
HAL
FIL
SBR
OAR
BRK
UAR
HAR
FIR
UHW
OBW
UBW
OLW
ULW
OSR
BEC
OSL
USR
USL
FUR
FBR
FUL
FBL
Head
SEH: Viewing Ray (0 DOF)
KO: Head (3 DOF)
Torso / Spine
OHW: Upper Cervical Spine (3 DOF)
BRK: Chest (1 DOF)
UHW: Lower Cervical Spine (3 DOF)
OBW: Upper Thoracic Spine (3 DOF)
UBW: Lower Thoracic Spine (3 DOF)
OLW: Upper Lumbar Spine (3 DOF)
ULW: Lower Lumbar Spine (2 DOF)
Right Arm
SBR: Right Shoulder Plate (2 DOF)
OAR: Right Upper Arm (3 DOF)
UAR: Right Forearm (2 DOF)
HAR: Right Hand (2 DOF)
FIR: Right Fingers (1 DOF)
Left Arm
SBL: Left Shoulder Plate (2 DOF)
OAL: Left Upper Arm (3 DOF)
UAL: Left Forearm (2 DOF)
HAL: Left Hand (2 DOF)
FIL: Left Fingers (1 DOF)
Pelvis (Hierarchy Origin)
BEC: Pelvis (6 DOF)
Right Leg
OSR: Right Thigh (3 DOF)
USR: Right Lower Leg (2 DOF)
FUR: Right Foot (2 DOF)
FBR: Right Foot Ball (1 DOF)
Left Leg
OSL: Left Thigh (3 DOF)
USL: Left Lower Leg (2 DOF)
FUL: Left Foot (2 DOF)
FBL: Left Foot Ball (1 DOF)

Figure 2. Anthropometric human model used by the motion tracker. **Left:** Inner model with 28 body joints. **Center:** Hierarchical structure of the inner model and its full parameterization by joint angles as provided in our data set. **Right:** Example instances of the parameterizable outer model. Note that joint angle representations may differ when performing the same action by differently sized models.

of them are computationally demanding and lack a proof of robustness necessary for long-term tracking. In recent years, a tendency towards learning based methods has been observed, to overcome the computational burden of searching the high-dimensional human pose space. While some systems try to infer a direct mapping from observed image features to articulated human poses [9, 6], others learn priors for human dynamics to provide a better prediction [19]. Both of these directions are unable to detect arbitrary, previously unobserved motions, making it difficult to apply them to the observation of everyday activities. The HumanEva [17] data sets provide a means to benchmark markerless motion capture systems by comparing tracking results with ground truth data. It consists of motions like walking, running or punching that are performed by several actors in an otherwise empty room without any interactions with objects or the environment.

A detailed survey on existing approaches for action recognition is provided by Krüger *et al.* [11].

3. The TUM Kitchen Data Set

The TUM Kitchen Data Set contains observations of several subjects setting a table in different ways (each recorded sequence between 1-2 min). All subjects perform more or less the same activity, including the same objects and similar locations. Variations include differently ordered actions, a different distribution of tasks between the left and the right hand, and different poses that result from different body sizes. Since the subjects are performing the actions in a natural way — apart from the sometimes unnatural transport of only one object at a time — there are fluent transitions between sub-actions, and actions are performed in parallel using both the left and the right hand.

Some subjects perform the activity like a robot would do, transporting the items one-by-one, other subjects behave more natural and grasp as many objects as they can at once. There are also a few sequences where the subject repetitively performed actions like picking up and putting down a cup (∼5 min each).

The data is publicly available for download at Researchers who are interested in the complete set of high-resolution image data are asked to contact the authors to coordinate shipping a hard disk or a set of DVDs. Please cite the paper at hand when publishing results obtained using the TUM Kitchen Data Set and send an email with the citation and a copy of your publication to the first author of this paper.

3.1. Data

The TUM Kitchen Data Set has been recorded in the TUM kitchen [4], a sensor-equipped intelligent kitchen environment used to perform research on assistive technologies for helping elderly or disabled people to improve their quality of life. The recorded (synchronized) data consists of calibrated videos, motion capture data and recordings from the sensor network, and is well-suited for evaluating approaches for motion tracking as well as for activity recognition. In particular, the following modalities are provided:

- Video data (25 Hz) from four static overhead cameras (Fig. 1), provided as 384x288 pixel RGB color image sequences (JPEG) or compressed video files (AVI). Additionally, full resolution (780x582 pixel) raw Bayer pattern video files are available.

- Motion capture data extracted from the videos using our markerless full-body motion tracking system [2]. Data is provided in the BVH file format[6], which contains the 6 DOF pose and the joint angle values as defined in Fig. 2. In addition to the joint angles, global joint positions are available as a comma-separated text file (CSV, one row per frame, the first row describes the column datatype). The data has been post-processed when necessary.

- RFID tag readings from three fixed readers embedded in the environment (sample rate 2Hz).

- Magnetic (reed) sensors detecting when a door or drawer is opened (sample rate 10Hz).

[6] http://www.cs.wisc.edu/graphics/Courses/cs-838-1999/Jeff/BVH.html

- Each frame has been labeled manually, as explained in Section 4.3. These labels are also provided as CSV files with one row per frame.

In addition, internal and external camera calibration parameters, as well as a more detailed technical documentation, are available from the web site.

3.2. Challenges

When performing action recognition on realistic mobile manipulation tasks, one notices several challenges which are hardly reflected in most of today's methods and data sets. In the following, we provide insights we gained from analyzing the activity data from different subjects:

Variation in performing activities: From the low motion level up to the sequence of actions, there are many degrees of freedom in how everyday activities can be performed (Fig. 3). This high variability needs to be taken into account when classifying human activities.

Continuous motions: People do not stop in between two motions, so there is no strict separation between single actions. To deal with this problem, sophisticated methods for motion segmentation are required, as actions can no longer be considered to be well-separated by default.

Parallelism: Humans often use both hands in parallel, sometimes even while walking at the same time (Fig. 3). Actions performed with the left and the right hand start at different times, and may overlap temporally and spatially. Sometimes actions are started with one and finished with the other hand, resulting in odd motion sequences per hand.

Difference in body sizes: Differences in body size may influence the way motions are realized. For instance, we found that tall subjects put objects on the table mainly by flexing the spine instead of the elbow joint.

Systems that are to operate in real environments have to cope with these challenges, and since the recognition rates of widely used, simpler data sets have become very good (improving only marginally), we believe that it is time to scale towards reality and more challenging data. To provide a gradual increase in complexity, we also provide sequences in our data set where the motions have been performed with clearer structure (e.g. objects are always grabbed with the same hand, less parallelism of actions).

4. Data Acquisition

In this section, we explain in more detail how we acquired the data. On top of the challenges for action recognition mentioned in the last section, the following requirements make the acquisition of such a comprehensive data set a difficult task:

Non-intrusive data acquisition: Apart from privacy issues, monitoring systems in a household scenario should be as unintrusive as possible. Attaching accelerometers or markers, as required by many motion capture systems, to real people in their everyday lives, or asking them to wear special skin-tight clothes is certainly not feasible. Therefore, the system has to perform all analysis based on sensors embedded in the environment and has to be able to track people independently of their body size or their clothing.

Interaction with the environment: Opening doors, picking up objects or being occluded by pieces of furniture causes significant changes of the human silhouette that may well confuse most of today's state-of-the-art markerless motion capture systems.

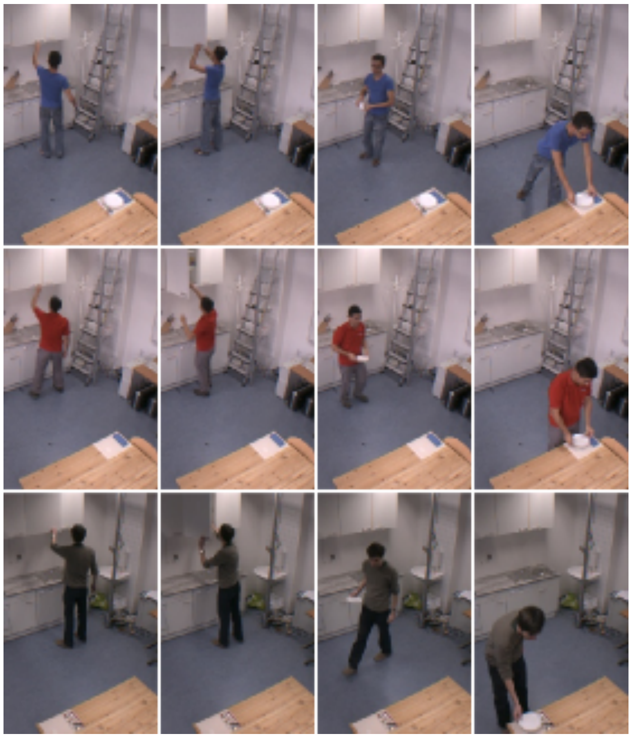

Figure 3. The same action performed by different subjects. Notice the use of different hands for the same action, the fluent transitions or parallelism of actions (e.g. opening cupboard and grabbing cup), and the difference in motion based on the size of the human.

4.1. Sensor-equipped Kitchen Environment

RFID and reed sensor data was recorded using the sensor network in the kitchen [4]. Tagged objects include the placemat, the napkin, the plate and the cup. Metallic items like pieces of silverware cannot be recognized by the RFID system, therefore, those object interactions are not recorded by the RFID system. Approximate positions of the RFID sensors and the cameras are given in Fig. 4 (left).

The subjects were asked to set the table according to the layout in Fig. 4 (right). Initially, the placemat and the napkin were placed at location A, the cup and plate in an overhead cupboard at location B. Silverware was in a drawer in between A and B. During the first recordings (the episodes are marked with a number starting with 0-*), the subjects

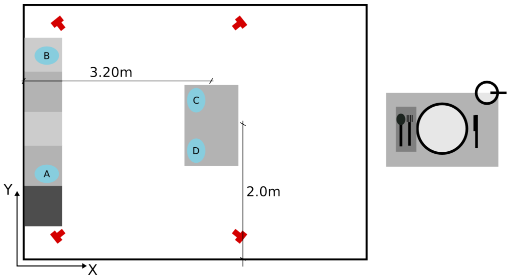

Figure 4. Spatial layout during the recording sessions. **Left:** Map of the kitchen with the approximate locations of the cameras (red) and the RFID sensors (blue ellipses). **Right:** Layout of the single items on the table.

placed the objects at location C, during the second set of recordings (episodes 1-*), they were put at location D.

4.2. Human Motion Tracking

Human motion capture data provides crucial information for understanding complex manipulation activities such as setting a table. The only realistic possibility to acquire such data to date has been the use of commercial marker-based tracking systems (optical, magnetic or inertial). However, using such systems in real scenarios is infeasible, as they are intrusive, expensive, difficult to set up, and often require a lot of post-processing. We have developed a markerless motion tracking system tailored towards application in everyday environments (Fig. 5). Our system comes with several improvements over both marker-based and state-of-the-art markerless motion capture systems:

- Setup is fairly easy, cheap and unintrusive, requiring only the placement of several cameras in the environment (three or more, depending on the amount of occlusions in the environment; see Fig. 1).
- We derive a full body pose for each timestep that is defined by an accurate 51 DOF articulated human model and corresponding estimated joint angles (Fig. 2 and 5). This also enables us to calculate the trajectories of specific body parts, e.g. hand trajectories during a pick and place action.
- The system is functional without a preceding training phase and is unconstrained with respect to the types of motions that can be tracked.
- By incorporating their appearance, we are able to track subjects that act and interact in realistic environments, e.g. by opening cupboards or picking up objects.
- The markerless tracker enables us to record realistic video sequences of human activities clear of any obstructive technical devices.

We will now briefly discuss the technical details of our system, and refer to [2] for a more detailed description. Our system estimates human poses in a recursive Bayesian framework using a variant of particle filtering. Each particle represents a human pose, given as a vector of joint angles. To account for the high dimensionality of the tracking problem, we developed a sampling strategy [3] that is a combination of partitioning of the parameter space [13] with a multi-layered search strategy derived from simulated annealing [8]. While the partitioning strategy enables us to take advantage of the hierarchical nature of the human body to reduce tracking complexity, the annealing strategy makes the system more robust to noisy measurements and allows us to use larger partitions at once that can be more accurately observed. We use a simple constant pose assumption with Gaussian diffusion to predict particles, and evaluate particle weights based on a comparison of projected and extracted silhouettes.

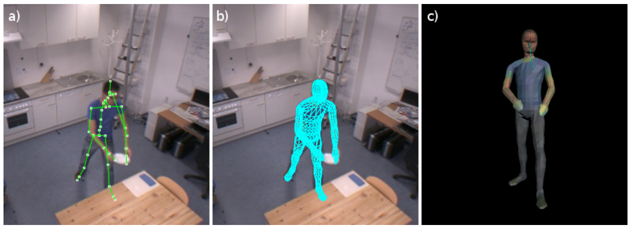

Figure 5. Human motion tracking (one of four cameras): a) inner model b) outer model c) virtual 3D view with appearance model.

Our method is able to track subjects performing everyday manipulation tasks in realistic environments, e.g. picking up objects from inside a kitchen cupboard and placing them on a table (Fig. 5). Dynamic parts in the environment (such as objects being manipulated or opening doors) are filtered and ignored when evaluating particle weights based on a comparison between expected background and known foreground (human) appearance when evaluating particle weights. Occlusions from static objects in the environment (e.g. tables) are dealt with by providing blocked regions that will only be evaluated in areas that resemble the learnt foreground (human). As a rule of thumb for occlusion handling, every body part should be visible by (at least) 2-3 cameras to achieve good tracking accuracy. Therefore, areas with heavy occlusions should be covered by more cameras to gather sufficient information for successful pose estimation.

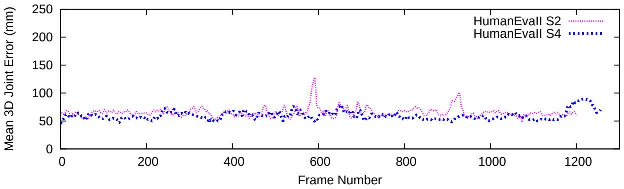

Figure 6. Accuracy and robustness of markerless motion capture as evaluated using the *HumanEva II* benchmark.

Validation of our system on the *HumanEva II* test suite [17] yielded mean Euclidean joint errors around 5-6 cm (Fig. 6), with error measurements partly corrupted due to a systematic error between relative joint positions in our model and the ground truth model. What is more important, the tracking accuracy stays approximately constant throughout the tested sequences, proving the robustness of our method. Videos of the performance on more complex scenarios, involving occlusions, dynamic environments and object manipulations, are available at `http://memoman.cs.tum.edu`.

We have used the presented tracker to acquire the motion capture data for our data set. Whenever tracking problems occurred (mostly while raising the right hand to reach towards the right outer cupboard, a motion insufficiently covered by our camera setup), we have manually post-processed the motion capture data to ensure high quality. While there are 84 DOF in the provided BVH files, the sequences were effectively tracked with 41 DOF. The difference is due to restricted DOFs in the human model (e.g. the elbow has only 2 DOF, see Fig. 2), an interpolation of the spine parameters during tracking, and the non-consideration of the hand/finger/foot parameters due to the insufficient resolution of the video data. The processing time of our method is around 20 sec/frame on common PC hardware, resulting in overall processing times per table-setting sequence of around 8 h.

4.3. Labeling

We manually labeled the data to provide a ground truth for motion segmentation and to allow for supervised training of segmentation algorithms. The left hand, the right hand and the trunk of the person are labeled separately to take the high degree of parallelism into account: Often, one hand is opening a cupboard door, while the other one is reaching towards a cup, and while the body is still moving towards the cupboard. Such parallel, overlapping actions are common in the data set. If a global label for the human action is desired, it can easily be constructed from the more detailed labeling.

When integrating the motion segmentation with higher levels of abstraction, it is important that the segments have well-defined semantics. We chose to represent each segment as an instance of a motion class in our action ontology (Fig. 7), which is inspired by the Cyc ontology [14]. Properties of the action, e.g. which object is being picked up or which cupboard is being opened, can easily be specified as properties of the respective instances.

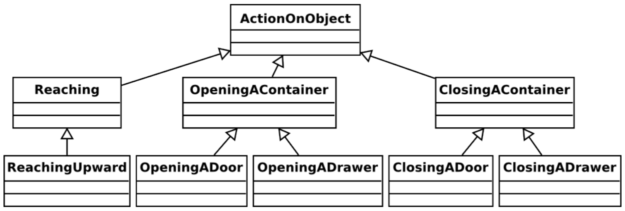

Figure 7. Excerpt from the action ontology used for labeling the data set. Special motions like *ReachingUpward* are defined as subclasses of the more general *Reaching* motion.

The set of labels describing the hand motion consists of *Reaching* towards an object or the handle of a cupboard or drawer, *TakingSomething*, *LoweringAnObject*, *ReleasingGraspOfSomething* after having put down an object or closed a door, *OpeningADoor*, *ClosingADoor*, *OpeningADrawer*, *ClosingADrawer*, and *CarryingWhileLocomoting*.

The last class describes all frames when the subject is carrying an item or, since the difference in the pose is often negligible, when the subject is moving from one place to another without carrying anything. It further serves as a catch-all class and contains the background motion between the actions. When acting naturally, humans perform surprisingly many irrelevant movements, such as reaching halfway towards an object, thinking about it, retracting the hands and doing something else first. If these motions are not long and articulated enough to be reliably recognized, they are put into the catch-all class, which therefore comprises rather diverse movements. The alternative, labeling every short, partial motion, would result in a large set of classes, ambiguities in the labeling, and hardly detectable short motion segments.

To take the very different postures for object interactions with the overhead cupboards into account, we added the classes *ReachingUpward* and *TakingSomethingFromAbove* which are specializations of the respective motion classes.

The motion of the person in the environment is described by the *trunk* label, which only has two possible values: *StandingStill* and *HumanWalkingProcess*, which indicates that the subject is moving.

5. Applications

We now show two use cases of the data set: inferring the labels from the motion capture and sensor data, and integrating the data into a knowledge processing system.

5.1. Motion Segmentation

This chapter is about our approach to automatically split and classify the motion capture data, i.e. to infer the labeled segments from the data. The system uses linear-chain Conditional Random Fields (CRF) (Fig. 8) to identify motion segments and represent them as instances of classes such as *Reaching* or *TakingSomething*.

As input, we use a large set of nominal features that can be split into two groups: The first group consists of pose-related features, which denote e.g. if the human is extending or retracting the hands, if the hands are expanded beyond a certain threshold, or if they are lifted above a certain limit. We also use discretized angles of the joints that are significant for the task at hand, like shoulder and elbow joints.

The second group of features includes information from the environment model and the sensor network, and indicates e.g. if the human is currently carrying an object, if a hand is near the handle of a cupboard, or if a drawer is being opened. Note, however, that all sensors only detect single events, while the labels are assigned to temporally extended motions, and that these events can be shifted due to different sample rates and a rather long detection range of e.g. the RFID sensors.

These features are combined, and CRFs are learned on a labeled training set. The CRF classifiers are integrated into the knowledge processing system and directly create a sequence of motion segments which are represented as instances of the respective semantic classes.

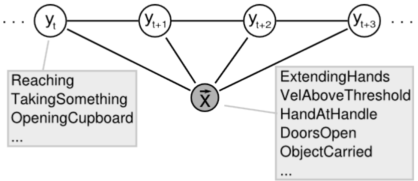

Figure 8. Structure of the CRF for segmenting human motions.

We evaluated our models using leave-one-out cross validation on the episodes *1-0* till *1-4*, i.e. the classifier has been trained on four sequences and tested on the remaining one. All sequences were of the same subject. The overall classification accuracy reaches 82.9% of the frames, or 93.1% on the best sequence. The combined confusion matrix, summing up the results of all test runs, is shown in Table 1. Bad results for the classes *TakingSomething* and *LoweringAnObject* result from the fact that these classes are only present in one of the episodes; so if the classifier is trained on the remaining episodes, it does not know this class at all and therefore must misclassify the respective frames.

We further tested the cross-subject classification accuracy by training on the episodes *1-0* till *1-4*, performed by subject *S0*, and tested on the episode *1-6*, performed by subject *S2*. These two subjects differ significantly in their body size and in the way how they move, which results in a lower accuracy of 62.77% of the frames.

	Reaching	ReachingUp	TakingSomething	LoweringAnObject	ReleasingGrasp	OpeningADoor	ClosingADoor	OpeningADrawer	ClosingADrawer	Carrying (Idle)
Reaching	152	25	7	0	3	0	0	45	6	64
ReachingUp	60	87	1	0	0	4	25	0	0	49
TakingSomething	0	0	0	0	24	0	0	45	2	0
LoweringAnObject	15	0	0	0	0	0	0	0	0	23
ReleasingGrasp	14	0	2	0	227	2	19	0	43	116
OpeningADoor	19	9	1	0	16	416	34	0	1	49
ClosingADoor	0	0	0	0	52	16	299	0	0	13
OpeningADrawer	27	11	41	0	64	0	0	312	39	6
ClosingADrawer	0	0	0	0	60	0	0	73	233	0
Carrying (Idle)	54	17	9	2	261	27	1	11	10	5717

Table 1. Confusion matrix of the classification of the left hand motion. Not all of the labels in the data set occur in the actions performed by the left hand.

While these are promising results, they also show that the classification of even seemingly simple everyday manipulation tasks can be very challenging, especially if they involve many classes and different subjects, and that more research is necessary to solve such problems.

5.2. Integration into a Knowledge Base

Since all motion segments are instances of well-defined classes in the ontology, it is possible to integrate them into a knowledge processing system in order to perform reasoning about the actions, their properties and parameters.

In our knowledge processing system, described in [18], this integration is solved using *computable* classes and properties. *Computable* relations load external information into the knowledge representation system, e.g. by sending a query to a database, or compute relations between instances, e.g. whether one object is on top of another. Here, computables are used to determine parameters of motion segments and actions, like an object being picked up or a cupboard being opened, which are determined by relating the segments to events simultaneously observed by the sensor network.

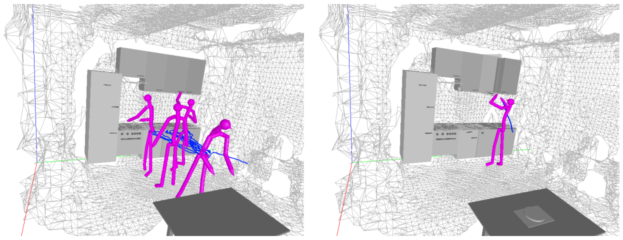

Figure 9. **Left:** Human pose sequence for setting a table. **Right:** Human pose sequence for taking a plate out of the cupboard.

An example of such queries is to select trajectories that serve a certain purpose. Fig. 9 visualizes the results of such queries: On the left is a query for all postures that are part of a tablesetting activity, on the right a query 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)
```

The *owl_query* predicates select the desired actions, namely all actions of type *TakingSomething* that are performed on an object of type *DinnerPlate* as a subevent of a *SetTable* activity. All poses that belong to the resulting actions are read with the *postureForAction* predicate. The pictures in this section are visualized results of queries to the knowledge processing system. Though the traces are only drawn for a single joint, the result includes the full human pose vectors for each point in time.

As described in [18], the data can also be used for learning the place from which an action is usually performed. Fig. 10 sums up the main steps: The observed positions

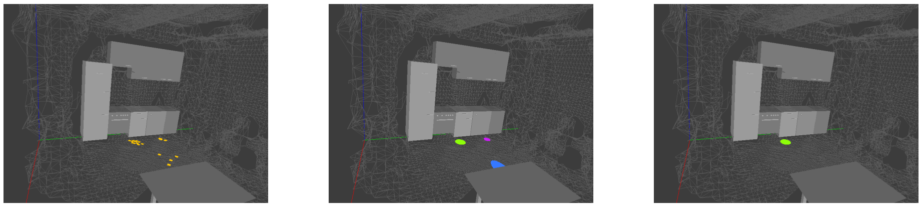

Figure 10. Learning places of interest. **Left:** Observed positions of manipulation actions. **Center:** The positions clustered into places. **Right:** The result that has been learned as the ”place for picking up pieces of tableware”.

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. 10 right) or to classify observations in order to find out about the most likely action performed from this place.

6. Conclusion

We presented the TUM Kitchen Data Set as a comprehensive resource for researchers in the areas of motion tracking, motion segmentation and action recognition. The data set already contains several multi-modal observations of table-setting activities performed by different subjects, and will be continuously extended over the course of the following months.

While the evaluation of different approaches on a common task is very important to judge their performance, it is also necessary to gradually increase the task complexity and move towards more realistic data as the systems improve. The TUM Kitchen Data Set is one step into this direction.

In addition to the description of the data, we also included experiments in which we used the TUM Kitchen Data Set, namely a system for segmenting the observed motion into the semantic classes given by the labeling, and a method for including the data into a formal knowledge representation.

7. 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] P. Azad, T. Asfour, and R. Dillmann. Toward an Unified Representation for Imitation of Human Motion on Humanoids. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2007.

[2] J. Bandouch and M. Beetz. Tracking humans interacting with the environment using efficient hierarchical sampling and layered observation models. IEEE Int. Workshop on Human-Computer Interaction. In conjunction with ICCV2009. To appear, 2009.

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

[4] M. Beetz et al. The assistive kitchen — a demonstration scenario for cognitive technical systems. In *IEEE 17th International Symposium on Robot and Human Interactive Communication (RO-MAN), Muenchen, Germany*, 2008.

[5] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *ICCV*, pages 1395–1402, 2005.

[6] L. Bo, C. Sminchisescu, A. Kanaujia, and D. Metaxas. Fast algorithms for large scale conditional 3d prediction. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, June 2008.

[7] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001.

[8] J. Deutscher and I. Reid. Articulated body motion capture by stochastic search. *International Journal of Computer Vision (IJCV)*, 61(2):185–205, 2005.

[9] K. Grauman, G. Shakhnarovich, and T. Darrell. Inferring 3d structure with a statistical image-based shape model. In *ICCV ’03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 641, Washington, DC, USA, 2003. IEEE Computer Society.

[10] R. Kehl and L. V. Gool. Markerless tracking of complex human motions from multiple views. *Computer Vision and Image Understanding (CVIU)*, 104(2):190–209, 2006.

[11] V. Krüger, D. Kragic, A. Ude, and C. Geib. The meaning of action: a review on action recognition and mapping. *Advanced Robotics*, 21(13):1473–1501, 2007.

[12] D. Kulić, W. Takano, and Y. Nakamura. Incremental learning, clustering and hierarchy formation of whole body motion patterns using adaptive hidden markov chains. *International Journal of Robotics Research*, 27(7):761–784, 2008.

[13] J. MacCormick and M. Isard. Partitioned sampling, articulated objects, and interface-quality hand tracking. In *ECCV ’00: Proceedings of the 6th European Conference on Computer Vision-Part II*, pages 3–19, London, UK, 2000. Springer-Verlag.

[14] C. Matuszek, J. Cabral, M. Witbrock, and J. DeOliveira. An introduction to the syntax and content of Cyc. *Proceedings of the 2006 AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering*, pages 44–49, 2006.

[15] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *17th International Conference on Pattern Recognition (ICPR 2004)*, volume 3, pages 32–36, Aug. 2004.

[16] Y. Sheikh, M. Sheikh, and M. Shah. Exploring the space of a human action. In *ICCV ’05: Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV’05) Volume 1*, pages 144–149. IEEE Computer Society, 2005.

[17] L. Sigal and M. J. Black. Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. Technical report, Brown University, 2006.

[18] M. Tenorth and M. Beetz. KnowRob — Knowledge Processing for Autonomous Personal Robots. IEEE/RSJ International Conference on Intelligent RObots and Systems. To appear., 2009.

[19] R. Urtasun, D. Fleet, and P. Fua. 3D People Tracking with Gaussian Process Dynamical Models. In *Conference on Computer Vision and Pattern Recognition*, pages 238–245, 2006.