Low Cost Activity Recognition Using Depth Cameras and Context Dependent Spatial Regions

(Extended Abstract)

Michael Karg
Institute for Advanced Study,
Technische Universität München
Lichtenbergstrasse 2a, D-85748 Garching
kargm@in.tum.de

Alexandra Kirsch
Department of Computer Science,
University of Tübingen
Sand 14, D-72076 Tübingen
alexandra.kirsch@uni-tuebingen.de

ABSTRACT

Recognition of human activities is usually based on expensive sensor setups to extract rich information such as body posture or object interaction. We investigate the use of inexpensive depth cameras to perform activity recognition using context dependent spatial regions with two different approaches: Spatio-Temporal Plan Representations and Hierarchical Hidden Markov Models. We evaluate both approaches in a simulated and a real-world environment.

Categories and Subject Descriptors

I.2.11 [**Distributed Artificial Intelligence**]: Intelligent agents; I.2.8 [**Problem Solving, Control Methods, and Search**]: Heuristic methods

General Terms

Experimentation, Human Factors, Algorithms

Keywords

Activity Recognition, Human Robot Interaction, Agents for improving human cooperative activities

1. ACTIVITY RECOGNITION

Activity recognition is a key capability for autonomous robots to interact with humans. Challenges include the high variability of human behavior, the representation of human activities consisting of several steps, but also the cost and acceptance for the necessary sensors. The goal of our work is to provide a household robot with information about ongoing human activities such as preparing a meal, so that the robot can proactively offer its help, monitor for failures or take care not to interfere with the human. Our activity recognition is based only on locations, which can be detected relatively reliably by state-of-the-art sensors, and the temporal order and duration in which they are visited. Additional object detection is discussed below.

We assume having a semantic map of the environment, specifically including furniture. From a dataset of 12 participants performing different pick-and-place tasks we learned where humans are generally located when performing actions (such as grasping) relative to reference objects (such as the table). The result is an extended spatial map of *Context Dependent Spatial Regions* with semantically annotated regions, at which pick-and-place actions are usually performed, relating each location with the furniture object on which the manipulated object stands. Once the regions are learned, they can be carried over to other environments as long as a semantic annotation of the furniture is available. From a map of Context Dependent Spatial Regions we generate two representations that serve as input to two activity recognition strategies:

Spatio-Temporal Plan Representations (STPRs) from our previous work [3], represent a human activity as a sequence of spatial regions visited throughout the task, together with the time spent at each location. Activity recognition is done by matching two such sequences. We use the *Generalized Levenshtein Similarity (GLS)* on a string representation of the sequences of locations as the similarity measures of an observed action (i.e. location) sequence with a known one.

Hierarchical Hidden Markov Models (HHMMs) as used by Bui et al. [2], are a hierarchical form of HMMs. On the lowest level they represent the regions with their transition probabilities, on the highest level are the human activities, also with transition probabilities. We use the Forward-Backward Algorithm to estimate the posterior marginals over all activities.

2. EVALUATION

For evaluation, we use data of the Human Morning Routine Dataset [4] with motion tracking data of a male person executing his morning routine over 14 days in simulation as well as in reality. We used the learned context dependent spatial regions applied to our experimental kitchen. To keep the effort limited, we manually generated STPRs and the state transitions of the HHMMs of the activities using sequences of context dependent spatial regions based on the spatial model. However, there are also approaches to learn such models from observations [3, 1].

For lack of other data, the following evaluations use the same dataset used to manually create the STPR and HHMM model. This is not ideal, but the variation in the data is large enough to pose a challenge. Assuming that such models

Appears in: *Alessio Lomuscio, Paul Scerri, Ana Bazzan, and Michael Huhns (eds.), Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2014), May 5-9, 2014, Paris, France.*

a) Simulated data

Activity	Prec. (%)	Recall (%)	Acc. (%)
Drink water	66.3	62.5	86.8
Prepare cereals	95.1	96.6	94.4
Prepare curd cheese	63.8	46.5	62.8
Clean table cereals	87.9	64.7	94.0
Clean table curd cheese	45.2	44.7	89.5
Prepare work	44.6	68.0	92.6

b) Real world data

Activity	Prec. (%)	Recall (%)	Acc. (%)
Drink water	35.9	37.0	76.4
Prepare cereals	51.9	67.5	62.9
Prepare curd cheese	34.8	25.0	63.0
Clean table cereals	68.4	23.2	82.3
Clean table curd cheese	85.8	34.1	84.9
Prepare work	63.4	91.3	92.6

Table 1: Average precision, recall and accuracy for 12 experiments of the simulated and real data using only locations with HHMMs.

could be learned with large data sets, the difference between learning and evaluation data would also be small.

Recognizing single activities.

In a first experiment, we examined how well single activities were recognized. Six activities were known, for example *prepare cereals*, and we used data from simulated runs as well as real-world data recorded with Kinect sensors.

Those runs show the disadvantage of STPRs representing a single action sequence. Variations in task execution are not represented, thus leading to high confusion rates. Still, in many cases, the correct action is recognized (even though it wins only closesly) or a similar action (for example, the system recognizes *prepare curd cheese* instead of *prepare cereals*). The results were worse than in previous work [3]. The reason is most likely the setup in the environments: while in [3] all locations were separate, the locations in this setup overlapped. This leads to additional uncertainty over the semantic label of spatial regions, which the STPR model cannot represent.

The HHMM model fares considerably better, recognizing the correct action with higher certainty. The fact that HHMMs do not represent the durations of actions or repeated occurrences of actions was no problem with our data set, but could be problematic in other cases.

With both methods, the recognition rates are similarly high with real data as with simulation data. This shows that the strategy of only using locations is well suited to real-world setups.

Recognizing several actions.

When several actions are performed, the additional challenge of recognizing action transitions occurs. To recognize activities in a sequence, we only use the HHMM model, as it was clearly better for single activity recognition and because it is well suited to represent another layer of more abstract activities. Again we used the data from our morning routine data set. Table 1 shows the results of these experiments and Figure 1 illustrates a specific recognition run.

The data show that our system can mostly recognize the activities, but not as reliably as other systems. This is not surprising, as our data is more parsimonious, but the achieved recognition rates are not enough for a reliable recognition that allows for appropriate robot behavior. However,

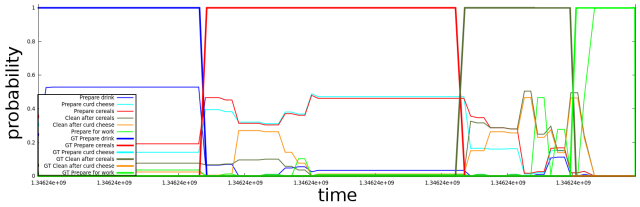

Figure 1: Probabilities of activity recognition for the morning routine from one of the 14 days in the real-world data set.

some of our activities are very similar, like the different food preparation activities. For a robot it may be sufficient to know that the user is preparing a meal, not necessarily which one. Therefore, we performed another experiment using the real data where we merged the food preparing activities and the cleaning activities into one activity each. We repeated the experiment and calculated precision, recall and accuracy, which resulted in improved accuracy values between 73.8 % and 94.4 % data.

We also investigated the possibility of increasing the accuracy by partly including object detections into the simulated experiment. Out of the 25 object interactions of the user, on average 15 were detected and the inclusion of those led to an increased accuracy between 91.0 % and 96.4 % for the simulated data. However, with real data, object detections were very unreliable, leading to accuracy values between 57.5 % and 91.0 % with an average of 76.41 %.

3. CONCLUSION

We presented an approach for activity recognition based on context dependent spatial regions in a kitchen environment using inexpensive depth cameras. We found that STPRs do not perform well in settings when regions are located very close to each other. HHMMs overcome this drawback and account for variations in task execution. By relying on locations as easily observable data, the uncertainty of the recognition lies more in the structure of the task than in the sensor noise.

4. REFERENCES

[1] M. Beetz, J. Bandouch, D. Jain, and M. Tenorth. Towards Automated Models of Activities of Daily Life. In *First International Symposium on Quality of Life Technology - Intelligent Systems for Better Living*, Pittsburgh, Pennsylvania USA, 2009.

[2] H.H. Bui, D.Q. Phung, and S. Venkatesh. Hierarchical hidden markov models with general state hierarchy. In *Proceedings of the National Conference on Artificial Intelligence*, pages 324–329. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004.

[3] M. Karg and A. Kirsch. Acquisition and Use of Transferable, Spatio-Temporal Plan Representations for Human-Robot Interaction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012.

[4] M. Karg and A. Kirsch. A human morning routine dataset. In *Proceedings of the Ninth International Conference on Autonomous Agents and Multiagent Systems (AAMAS), Extended Abstract*, 2014.