Acquisition and Use of Transferable, Spatio-Temporal Plan Representations for Human-Robot Interaction

Michael Karg[1], Alexandra Kirsch[2]
{kargm, kirsch}@in.tum.de

***Abstract*— Service robots that work together with humans in domestic and constantly changing environments should have a general understanding about their human partners and the tasks that are to be performed. This would enable them to verify their beliefs about the common tasks and the goals of their human partners and detect unexpected events and failures. In this paper we present a way of acquiring general, spatio-temporal plan representations from human motion tracking data in different environments. Using an annotated data set for table setting tasks in a typical kitchen environment, we first cluster the static positions of the participants and create a spatial model relative to furniture objects that are given by a semantic map and linked to a knowledge base. Based on this spatial model learned in one kitchen, we automatically generate spatio-temporal plan representations in different kitchen environments with known semantic maps. We show that our models can successfully be used to give a robot a basic understanding about a task executed by a human in three different environments. We evaluate the quality of our automatic generation of the plan representations and present an example application of plan supervision using a learned model from one kitchen to differentiate tasks performed by humans in other kitchens.**

I. Introduction

With service robots closely working together with humans in domestic environments, knowledge about the user becomes an inevitable part of the robot's knowledge base. When working cooperatively with a human partner, it would be beneficial for a robot to have models about the human and the tasks that are being executed in cooperation. Using such models, a service robot could be able to differentiate between nominal and exceptional behavior and use expectations about its human partner to detect and prevent unexpected events and failures. This would enable a robot to react adequately to different actions of its partner, thus being a more convenient and reliable helper to the human.

An intelligent robot should not be limited to one single environment. It therefore needs the ability to transform knowledge about plans to different locations and adapt to changes. Humans can adapt their knowledge and experiences to new situations after observing a single instance of a new concept and refine it incrementally [7]. So it is advantageous to generate general, transferable models of human behavior that a robot can use in all kinds of different environments.

With the support of the Technische Universität München - Institute for Advanced Study, funded by the German Excellence Initiative.

[1]Institute for Advanced Study, Technische Universitt München, Lichtenbergstrasse 2a, D-85748 Garching, Germany http://hcai.in.tum.de

[2]Department of Computer Science, University of Tübingen, Sand 14, D-72076 Tübingen, Germany http://hcai.in.tum.de

Our approach is based on the observation and recognition of human tasks and thus dependent on tracking human motion. Although there exists a variety of different human tracking systems, it is mostly time-consuming and expensive to equip the environment with sensors, which have to be set up, calibrated, synchronized, etc.. Recent development in game consoles led to cheap and easy to use depth-sensors that can also be used for human tracking, but is limited to a small area and is also limited in accuracy. The transferable spatio-temporal plan descriptions we present in this paper enable a robot to perform plan monitoring in different environments with an inexpensive sensor such as the Kinect.

If humans think about plans like setting a table, they do not have coordinates in their mind, but rather a more abstract, semantic comprehension about the actions it takes to set a table. Instead of x,y,z coordinates, humans see places in terms of relative locations to entities in the environment [1]. A human might for example rather say: "Get a plate from a cupboard and put it onto the table." instead of "Go to point 10,16,0 and move your hand to point 14,18,0.7, then ...". A robot's understanding of a location often depends on specific coordinates that are stored in reference to a specific map and are thus of limited use in other environments and fragile to changes. In this work we aim to decouple the models from the environment by using semantic maps to represent locations of a human in reference to objects that are related to his currently active task. These semantically annotated locations can then again be used to automatically generate transferable, spatio-temporal plan representations that are closely related to the way humans would describe a plan. In the example of a table-setting plan, this may look like this: "First go to the oven, second go to the table, third, ...". This way we obtain simple and general models of plans that give a robot an understanding of locations that is more human-like and as Kennedy et al. state, "... a system that uses representations and processes or algorithms similar to a person will be able to collaborate with a person better than a computational system that does not." [6].

As we will show in this paper, such plan representations can be transferred to other environments that have a semantic map representation and by also including information about the durations that a human spends at the semantically referenced locations, a robot can use those models for plan supervision in different environments and differentiate between tasks performed by the human.

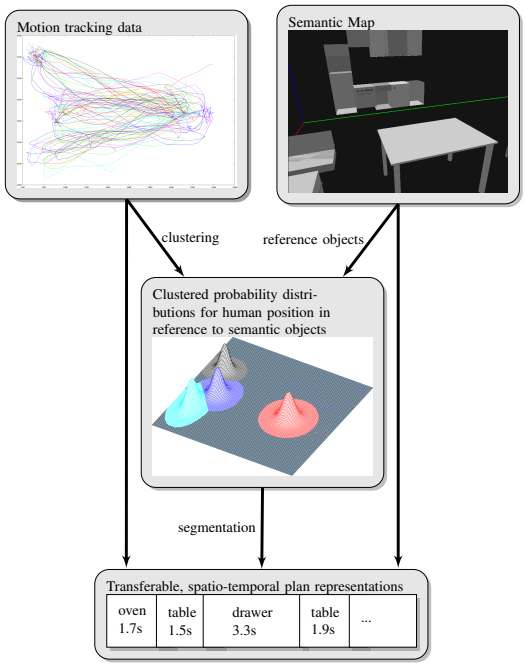

Fig. 1. An overview of our generation of general models about human locations during plan execution. Labeled motion tracking data is clustered for locations where the human interacts with objects. The clustered locations are then referenced to furniture objects obtained from a semantic map of the environment. This model is then used to do an automatic segmentation of the motion tracking data and create symbolic, time-line based models that are transferable to different environments.

Figure 1 shows an overview of our approach to generate the spatio-temporal plan representations. We perform a clustering of motion tracking data to generate probability distributions and use knowledge from a semantic map to semantically annotate the probability distributions to generate a spatial model. Based on this spatial model, we perform an automatic segmentation and create transferable, spatio-temporal plan representations.

In the remainder of this paper we will first present related work in this field, then explain how a simple model for a table setting task is generated. We then apply this model in different environments, give an example application of use and conclude at the end.

II. Related Work

Recent development in autonomous semantic mapping allows robots to have semantically annotated knowledge about their environment by autonomously generating semantic maps using laser scans, 3D-pointclouds and cameras. While some approaches focus on detecting room categories [21], others include handle detection of cupboards and drawers and the learning of articulation models for the opening and closing of the drawers and cupboards [4]. Tenorth et al. [18] offer a way to create knowledge-linked semantic object maps that combine information of semantic maps with common sense knowledge of publicly available data bases. This enables a robot to do reasoning about the objects in its environment, for example query its database for common locations of objects in its specific environment.

Humans generally tend to represent spatial regions not only geometrically but also according to their functional use. For a robot interacting in a human-populated environment, it must understand its environment in terms of human spatial concepts [21]. One step towards this understanding for machines is done by Liao et al. [8]. They use hierarchical conditional random fields to learn patterns of human behavior from GPS traces, recognize significant places that the human visits during everyday activities and label them according to their function (office, home, ...). They even transfer models learned from one person to another and estimate the test-person's intentions. Stulp et al. [15] propose a representation of the utility of positions in the context of action-related mobile manipulation. They define so-called *ARPlaces* as probability distributions in reference to the pose of objects to model the probability for a successful grasp. Klenk et al. [7] find that "the ability to understand and reason about spatial regions is essential for cognitive systems performing tasks for humans in everyday environments". In their work, they define context dependent spatial regions for cognitive systems that are learned by qualitative spatial representations and semantic labels. Also they identify similar environments and show that they are able to transfer their context dependent regions to them. Bennewitz et al. [3] state that for mobile robots working together with humans, knowledge about the locations of people in the environment is important and they use typical human motion patterns to actively maintain beliefs about positions of humans in the environment and their intentions. As in our approach, they assume that people performing everyday activities are not in permanent movement, but move between "resting places" while in contrast to our approach, they focus on the trajectories between the "resting places" and utilize Hidden Markov Models to estimate positions of the human. Another approach for human activity recognition that is based on Hierarchical Hidden Markov Models is proposed by Nguyen et al. [10], [11]. They use sequences of manually annotated locations to recognize and monitor high-level behaviors in an office environment. There are also approaches that perform activity recognition based on sequences of objects that the human interacts with. Buettner et al. [5] for example use RFID sensors with accelerometers and a kitchen equipped with antennas to show that they can successfully detect a wide range of human activities based on a sequence of objects that the human interacts with. Perkowitz et al. [13] use a similar RFID-based setup to detect human everyday activities that have been mined from the web using a Monte-Carlo based framework based on the detection of object-sequences. While the RFID-based approaches allow for the detection of a wide range of activities, they need many objects equipped with sensors and the environment and/or the human has to wear RFID readers which might not always be feasible in everyday situations.

Spatial regions can also be used to model nominal behavior of human plans as in Orkin et al. [12]. Here the authors automatically generate plans that represent nominal behavior of humans visiting a restaurant from human game play in a

computer game simulating a restaurant. They use the learned nominal models to generate and validate expectations and detect exceptional situations of the players.

III. Spatial Model Generation

A. The TUM Kitchen Dataset

For the generation of our spatial model, we use already available labeled real world motion-tracking data of the TUM kitchen dataset[1] [16] which offers several recordings of humans performing a table-setting task in a typical kitchen environment. The recordings consist of video sequences from four cameras, full body motion tracking data of the human, RFID tag readings and magnetic sensor readings from objects and the environment. The data has manually been labeled to provide a ground truth for motion segmentation and includes labels for the actions of the human body in general and each hand separately. The table setting scenario consists of 6 objects (placemat, napkin, fork, knife, plate, cup), that are stored at three different locations (drawer, cupboard, stove), and a table that is to be set. Each of the 10 participants executed one task where they where told to set a table in a way we will call *impaired-person* in the rest of the paper. Here they were only allowed to transport one item at a time. Two of the 10 test persons also executed a second table-setting task in a mode we call *able-bodied-person*, where they were allowed to carry several items simultaneously. For our approach we use the full body motion tracking data and the labels of the body and the hands to create a semantic model about human positions while performing a table setting task. We also use a semantic map of the environment [4] that is linked to the KnowRob knowledge base [18], [17] and provides us with information about objects in the environment such as cupboards and drawers.

B. Clustering static positions and assigning semantic reference points

Based on the assumption that a human is mostly not moving while interacting with objects, we extract positions from the TUM kitchen dataset where object interactions occur and the human is standing still. We therefore use the motion tracking data and the labels and account for pick and place actions and the opening or closing of cupboards or drawers. This step is done for the 10 impaired-person experiments and, using WEKA [19], we perform a clustering of the extracted human positions based on the Expectation Maximization Algorithm [9] that gives us the four clusters as shown in Figure 2. Here we assume to know the number of clusters that relate to the different storage locations of the objects involved in the plan. In the case of the TUM kitchen dataset, the objects are stored on the oven, in the drawer and in one of the cupboards at plan start, and when the plan is terminated, all of the objects are stored on the table at the place where the human wants to have breakfast. For every cluster, we fit a probability distribution in the global coordinate system to describe different positions

[1]http://ias.cs.tum.edu/software/kitchen-activity-data

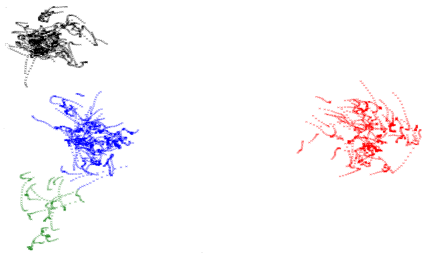

Fig. 2. Clustered locations of a human when interacting with objects during a table setting task in a kitchen.

the human visited during plan execution. Here we decided to use two dimensional Gaussian probability distributions since fitting Gaussians does not require a big effort and the clusters seemed to be almost normally distributed. However one could also think about using more elaborate probability distributions like the *ARPlaces* in the work of Stulp et al. [15].

But to build a general model that can be transferred to different kitchens, we need the human positions relative to the locations of storage places of the objects the human interacts on. This generalizes the model of the table setting task to every kitchen that has a semantic map. To achieve this independence of a specific coordinate frame, we use the semantic map of the environment and relate the means of the two dimensional Gaussians of the human locations to the positions of the storage locations of all objects. We assume that the position that a human chooses for interacting with objects depends on the opening direction of a container such as a cupboard or a drawer, if the object is stored in a container. In these cases, he first has to open the container before being able to grasp the object. So to generate meaningful semantic reference coordinate frames for the two dimensional Gaussians, we distinguish between two groups: objects that are stored *in* a container (e.g. a cupboard or drawer) and objects that are located *on* a piece of furniture.

1) Objects in containers: For containers like cupboards and drawers, we obtain the location of their centroid, the depth, width and height from the semantic map. We can also extract information about the location of the side where the door is located and the positions of the handle and the hinge of cupboards. We decided to use the middle point on the edge of the door of the container as origin for the reference coordinate frame as illustrated in Figure 3 on the left (assuming a 2-dimensional view of the environment). This position can easily be calculated given the information from the semantic map. The position of the hinge of a cupboard is used to decide where the y-axis of the reference coordinate frame is pointing to since usually the position of a human opening a door depends on which direction the door opens.

2) Objects on pieces of furniture: General surfaces have to be treated differently since they may not have a general orientation like the opening side of cupboards or drawers. In our table setting dataset this case happens for the placemat and the napkin which are located on the stove and on the

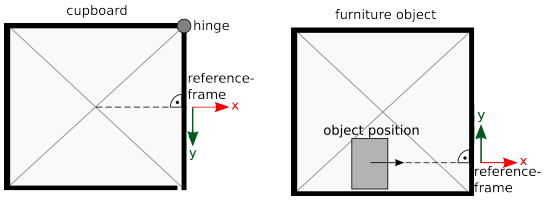

Fig. 3. The left figure shows how we calculate reference frames for human locations using information about a cupboard from a semantic map. The y-axis of the coordinate system of the reference points away from the hinge of the door. The right picture illustrates the calculation of the reference coordinate frames for objects that are located on pieces of furniture objects.

table. We take the position of the object on the furniture piece into account, because the position of the human varies depending for example on where on the table he wants to have breakfast. Since the data of the TUM kitchen dataset does not include the positions of all the objects that are used, we approximate their locations by using the full body motion tracking data to extract the locations of the hands of the human when pick and place actions occur (assuming consistent placing in all experiments). Due to the separate labeling of left hand, right hand and trunk, this can easily be done and averaging over the 10 experiments we obtain the approximate locations of all of the objects. We also calculate a 2D-orientation of the approximate object positions by drawing a vector from the approximated object location to the position of the human when a pick and place action occurs. Using the position and orientation of the object, we use the edge of the supporting furniture piece that the orientation vector points to as reference point as illustrated in Figure 3 on the right.

Now that we have defined reference coordinate frames for all of the objects involved in the table-setting task, we can put the clustered Gaussians from the motion tracking data into reference to the nearest furniture object involved in the human plan. This way we obtain our spatial model ψ as a set of locations that consist of Gaussians P_i linked to semantically annotated instances of furniture objects o_i in our semantic map:

$$\psi = \{l_1, l_2, ..., l_n\} \text{ with } l_i = (P_i, o_i)$$

So we linked observed positions of a human performing a table-setting task with furniture objects that have a semantic representation in the semantic map of the environment. We can now use this model to do a segmentation of the motion-tracking data of the TUM kitchen dataset or other motion tracking data to automatically generate symbolic task descriptions that also incorporate time.

IV. Spatio-Temporal Plan Descriptions

A. Concepts

We define a spatio-temporal plan description p_n as a sequence of n tuples that have a location l_i and a duration t_i as elements.

$$p_n = ((l_1, t_1), (l_2, t_2)..., (l_n, t_n))$$

An example of a spatio-temporal plan description of a table setting task can be see in Figure 4. The advantage of using locations in our plan representations instead of actions is that the robot does not have to do action recognition, which is a complex task and often not feasible for a robot that has to cope with problems such as self-occlusions of the human, limited sensor range, etc. The locations are defined by a set of probability distributions in the spatial model as described in section III-B and they are linked to the semantic map of the environment and thus to our knowledge base. This description allows us to compare different spatio-temporal plan representations in two different ways: the durations a human spends at different types of locations, and the plan patterns.

Durations: For the durations t_i, we define a confidence value that expresses how well the durations of the observed locations of the human fit to the model of a plan like our table setting task. We model the durations a human spends at every type of furniture object of our reference plan using Gaussian distributions that have been learned from observations (as we will show in section VI):

$$\varphi(t_i)_{l_i} = \frac{1}{\sigma_{l_i}\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{t_i - \mu_{l_i}}{\sigma_{l_i}}\right)^2\right).$$

$\varphi(t)_{l_i}$ defines the durations at the specific location as Gaussian probability distribution where μ_{l_i} and σ_{l_i} are the mean and the variance in the durations at location l_i of our reference-plan. To obtain a confidence measure c_p that describes how well the observed locations of the human fit to the model, we calculate the average confidence values over all durations t_i of the observed locations l_i for a plan p_n as follows:

$$c_p = \frac{\sum_{i=0}^{n} \frac{\varphi(t_i)_{l_i}}{\varphi(\mu_{l_i})_{l_i}}}{n}.$$

Plan patterns: In addition to the durations at specific locations, the pattern of a plan i.e. the sequence of the visited locations is also a significant (maybe even stronger) feature that describes a plan. To compare plan patterns without regard for the durations, we generate a string representation of our spatio-temporal plan representation. Using an acronym for the locations based on the objects as in Figure 5, the string representation of the plan shown in Figure 4 looks like this: "ADADCDBDBDBDCD". This way we can use string comparison methods, in our case a normalized Levenshtein distance metric called *Generalized Levenshtein Similarity (GLS)* [20], to compare the model of the table-setting task learned from the TUM kitchen dataset with models generated from observations of other tasks. The GLS is a string matching technique that is based on how many editing steps it takes to transform one string into another. A value of 1 defines a perfect match of the strings, while a value of 0 expresses no correlation at all.

B. Generation of Spatio-Temporal Plan Descriptions

Using our spatial model from section III-B, we can now segment the motion tracking data of the TUM kitchen dataset

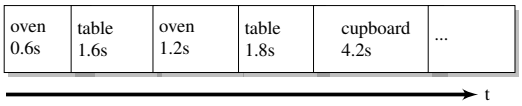

Fig. 4. Automatically generated, spatio-temporal plan representation of a table-setting task.

to obtain spatio-temporal plan representations of a table-setting task. To this end we examine the locations where the human is standing still for a short time and assign to it the semantic location with the highest probability according to our two dimensional Gaussian probability distributions. We consider the human "standing still" at locations where the center of mass of the human is not moving more than 25 cm within 0.5 seconds (0.5 m/s). We also take into account how long the human is staying in each of the locations and as a result of this segmentation, we obtain ordered, symbolic representations that we illustrate in the form of a timeline as shown in Figure 4. We performed this segmentation using the data of the TUM kitchen dataset and to evaluate our segmentation, we compare our symbolic plans with the ground truth. Out of the 10 experiments, 7 experiments have all locations correctly assigned and the symbolic plans correspond to the ground truth. In three of the experiments, the segmentation fails to recognize one instance of the human standing in front of the oven or the table. The reason for that error is that while grasping or releasing objects on general surfaces, the human sometimes is not standing still for sufficiently long, but rather grasps the object almost while moving. Nevertheless, except for these missing instances, all other locations are also correctly assigned and compared with the ground truth we achieve a GLS (as described in IV-A) of 97,15 % on average for all instances of our automatically generated spatio-temporal plan representations. A possible solution for the problem of the missing instances could be the detection of motion sequences where the human moves towards a furniture object and moves away in the opposite direction after a small amount of time which could be a clue for a pick and place task.

V. Transferring the Semantic Models to Other Kitchens

To test the transferability of our model to other environments, we defined two more experimental setups. For the first (*setup 1*), we used the same kitchen as used for the TUM kitchen dataset, but we varied the cupboard where the plates are stored in and also moved the table that is to be set by the human to another location. For the second experimental setup (*setup 2*), we used a completely different kitchen where the storage location for the plate and cup was a second drawer instead of a cupboard. Providing the new locations of the table and the knowledge about the storage locations of the other objects from a semantic map, the model can directly be applied to the new environment as shown in Figure 5.

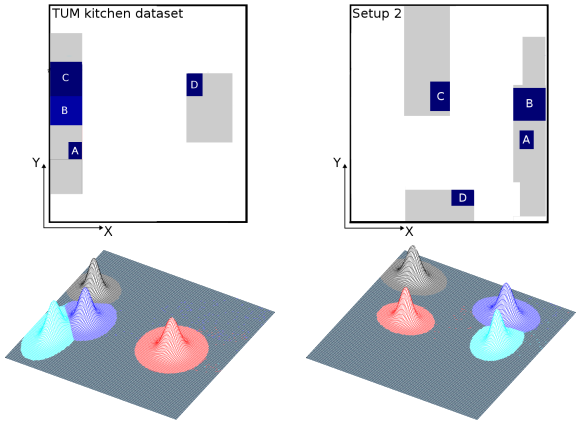

Fig. 5. The left picture shows the setup of the TUM kitchen dataset and the expected locations of a human while setting a table modeled as two dimensional Gaussians. Here A corresponds to the location of the placemat and napkin, B describes the drawer where the cutlery is stored, C represents the cupboard where the plate and cup are in and D is the place where the table is to be set. On the right picture, setup 2 is shown. Here our model is applied to a completely different kitchen. As illustrated on the lower right picture, the expected locations of our model have been automatically adapted to the new environment given that the locations of the objects storage places are known.

Using a Kinect for human motion tracking we recorded a new dataset of 8 persons in setup 1 and 10 persons in setup 2. The participants were researchers from the field of computer vision and robotics and did not have any idea about what the experiment would be about. They were told to perform a impaired-person table-setting task and an able-bodied-person table setting task as in the TUM kitchen dataset. We applied the same segmentation as in section IV-B and for setup 1, seven out of eight spatio-temporal plan descriptions were generated completely correctly. For setup 2, six out of ten spatio-temporal plan descriptions were generated completely correctly. As in section IV-B, the problems for the segmentation mostly result from the participants not standing still for long enough when grasping an object from a surface. Using the GLS as described in section IV-A, we calculate the similarity between our automatically segmented data and the ground truth and achieve a GLS of 98.22 % for setup 1 and 94.28 % for setup 2. These values indicate that automatic segmentation performs reasonably well in environments that differ from the original environment of the learned model and our spatial model can be transferred. We also compare the durations the human spends on average at different furniture objects to find out, if they are comparable in the different experiments. We averaged over the 10 experiments of the TUM kitchen dataset and the experiments in setup 1 and setup 2 and calculated how much time the human generally spends at one specific location. The means and variances for the TUM kitchen dataset and both experiments are shown in Figure 6.

In setup 1, the cupboard was located at the maximum tracking range of the Kinect sensor, which caused a high amount of sensor noise. This resulted in the tracked human *"jumping around"* while the real human was actually standing still in front of the cupboard. This is a reason why the

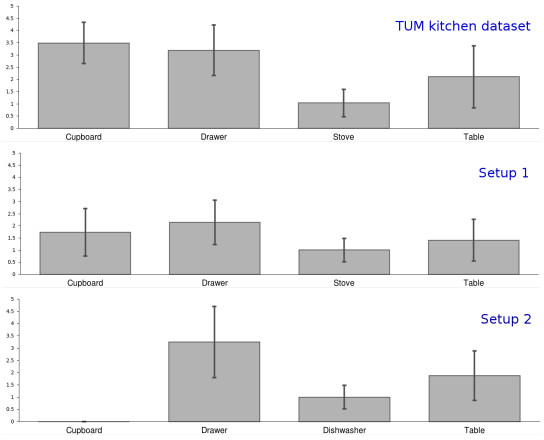

Fig. 6. The pictures show average durations of a human standing still at the locations of our model during a task execution with their means and variances. The upper picture shows the data from the TUM kitchen dataset, the middle is showing data from setup 1 and the lower picture represents setup 2.

durations of the human at the cupboard in setup 1 differs quite a lot from the TUM kitchen data. The other durations are similar for the different types of furniture objects. The participants in setup 2 had similar durations when performing pick and place tasks as in the TUM kitchen dataset, while the participants of setup 1 were a little faster at every location. So the durations in front of different types of locations depend on individual humans and should not be used as a strong evidence for pick and place actions, but since a tendency is clearly visible, they could still be used to give the robot a clue about the plan the human is executing as we will show in section VI-.1.

VI. Applications

The spatio-temporal plan representations that we explained in the last sections offer various possibilities of application for human robot interaction and can be applied in every environment for which a semantic map representation is available. Here we show an example application of our models.

Passive Plan Supervision

Using our model to perform a passive plan supervision using the durations and the patterns of our spatio-temporal plan representations to calculate confidence values that express how certain the robot is that the human has performed a specific plan.

1) Durations: As stated in section V, the durations that a human spends at certain locations differ from human to human, but we still wished to figure out if only information about the durations can enable us to distinguish between different plans, thus being able to identify typical pick and place tasks. Assuming that the durations at each location depend on the amount of manipulation, we recorded an additional experiment with the participants of our first and second experiment in which they were told to perform a cleaning task. The goal was to remove used dishes from the table and clean the dishes and the kitchenette. We chose this task because it uses similar locations as the table setting task and thus cannot be easily distinguished. For the generation of the reference plan, we use 10 experiments of the impaired-person table setting task from the TUM kitchen dataset as described in section III and IV. To test if we can successfully identify the impaired-person table setting task, we calculated confidence values as illustrated in section IV-A for the three different observed plans of every participant and for every experiment of our two experimental setups. Averaging over all experiments, we obtain confidence values according to the following table:

Task	c_p Setup 1	c_p Setup 2
Impaired-person table setting	0.524	0.593
Able-bodied-person table setting	0.448	0.506
Cleaning task:	0.191	0.350

We can see here that the robot is more confident of seeing a table setting task if a table-setting task was indeed performed. Even the able-bodied-person table setting tasks have a relatively high confidence value compared to the cleaning tasks. This can be explained by the fact that the time, a human spends at a storage location does not vary significantly if he picks up one single object or several objects. Although the durations may not be a strong indicator, they still enable a robot to distinguish between typical pick and place tasks such as the table-setting tasks and other tasks (e.g. cleaning) based only on the times the human spends at certain locations.

2) Plan patterns: Comparing the plan-patterns of all of our experiments using the GLS as explained in section IV-A, we compute the following values:

Task	GLS Setup 1	GLS Setup 2
Impaired-person table setting	0.982	0.943
Able-bodied-person table setting	0.429	0.429
Cleaning task:	0.357	0.340

The values indicate that in our experiments consisting of three different tasks, we are able to identify the table-setting task from our automatically segmented motion tracking data. In contrast to the durations, here the values of the impaired-person table setting task strongly deviate from the able-bodied-person table setting task since their patterns are quite different. A combination of durations and plan patterns can thus be used as a reliable classification scheme. While the durations alone can distinguish several tasks from each other, they fail when it comes to different modes of the same task. Those difference modes can be distinguished using the plan patterns, so for future work we will investigate combination methods for both measures to improve out task classification.

VII. Discussion

The experiments show that only using the sequence of locations is sufficient to distinguish several tasks from each other. This has the advantage that it can be done with low-cost sensors given a semantic map of the environment. However, there might also be some tasks that look similar with regards to the sequence of locations a human visits and

the durations at the specific places. In this case our models would not provide enough information to identify one single task, but we would still be able to identify several activities that are more likely to be executed by the human than others. To account for similar looking tasks, one could also use object detections at different locations which would introduce better distinguishability between the plan representations. Maybe even one single detection of a significant object used by the human could enable us to distinguish between tasks that would look similar when only using locations. Also a richer modeling of the human activities could help to increase the distinguishability and thus the recognition rates of activities that look similar. One example for rich human activity models are graph-like structures as introduced by Sridhar et al. in [14]. Also the inclusion of partial ordering constraints and hierarchies into the human activity models as presented by Beetz et al. in [2] could be a possible extension. A modeling technique like this could improve robustness in terms of different variations of tasks. For the generation of our spatial model, we set the position of the human in relation to the storage locations based on the direction to which the container opens or the position of an object on surfaces. But the relative position of the human can also depend on other factors like the location where he is coming from and the location where he will go next. To keep our state-space limited, we decided to neglect these effects for now, but for future work, the segmentation and the recognition rates in different environments could be improved by considering these effects.

VIII. Conclusion

This paper presented an approach for the generation of general, transferable spatio-temporal plan representations for human-robot interaction that uses human motion tracking data and semantic maps of the environment. We created a general model of a human performing a table-setting task and showed that this model can be used to perform an automatic segmentation of motion tracking data in different environments given a semantic map. We evaluated our segmentation and presented an example application of use in a basic plan supervision framework where we used only a Kinect sensor for motion tracking. We successfully distinguished a table-setting task from a cleaning task in two different environments. For future work, we plan to record more motion tracking data of humans performing different tasks in various environments and examine if our model scales to a variety of tasks and environments. We also plan to use the models in a real-time system for the detection of abnormalities during human plan execution.

References

[1] A. Aydemir, K. Sjoo, J. Folkesson, A. Pronobis, and P. Jensfelt. Search in the real world: Active visual object search based on spatial relations. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 2818–2824. IEEE, 2011.

[2] Michael Beetz, Jan Bandouch, Dominik Jain, and Moritz 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.

[3] M. Bennewitz, J. Pastrana, and W. Burgard. Active localization of people with a mobile robot based on learned motion behaviors. 2008.

[4] Nico Blodow, Lucian Cosmin Goron, Zoltan-Csaba Marton, Dejan Pangercic, Thomas Rühr, Moritz Tenorth, and Michael Beetz. Autonomous semantic mapping for robots performing everyday manipulation tasks in kitchen environments. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, San Francisco, CA, USA, September, 25–30 2011. Accepted for publication.

[5] M. Buettner, R. Prasad, M. Philipose, and D. Wetherall. Recognizing daily activities with rfid-based sensors. In *Proceedings of the 11th international conference on Ubiquitous computing*, pages 51–60. ACM, 2009.

[6] W.G. Kennedy, M.D. Bugajska, M. Marge, W. Adams, B.R. Fransen, D. Perzanowski, A.C. Schultz, and J.G. Trafton. Spatial representation and reasoning for human-robot collaboration. In *PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE*, volume 22, page 1554. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2007.

[7] Matthew Klenk, Nick Hawes, and Kate Lockwood. Representing and reasoning about spatial regions defined by context. In *AAAI Fall 2011 Symposium on Advances in Cognitive Systems*, 2011.

[8] Lin Liao, Dieter Fox, and Henry Kautz. Extracting places and activities from gps traces using hierarchical conditional random fields. *Int. J. Rob. Res.*, 26(1):119–134, 2007.

[9] G.J. McLachlan and T. Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley and Sons, 2008.

[10] Nam T. Nguyen, Dinh Q. Phung, Svetha Venkatesh, and Hung Bui. Learning and detecting activities from movement trajectories using the hierarchical hidden markov models. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2*, pages 955–960, Washington, DC, USA, 2005. IEEE Computer Society.

[11] N.T. Nguyen, H.H. Bui, S. Venkatsh, and G. West. Recognizing and monitoring high-level behaviors in complex spatial environments. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–620. IEEE, 2003.

[12] Jeff Orkin and Deb Roy. The restaurant game: Learning social behavior and language from thousands of players online. *Journal of Game Development (JOGD)*, 3(1):39–60, December 2007.

[13] Mike Perkowitz, Matthai Philipose, Kenneth Fishkin, and Donald J. Patterson. Mining models of human activities from the web. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 573–582. ACM, 2004.

[14] M. Sridhar, A.G. Cohn, and D.C. Hogg. Unsupervised learning of event classes from video. In *Proc. AAAI*, pages 1631–1638. AAAI Press. Menlo Park, 2010.

[15] Freek Stulp, Andreas Fedrizzi, and Michael Beetz. Action-related place-based mobile manipulation. In *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)*, 2009.

[16] Moritz Tenorth, Jan Bandouch, and Michael Beetz. The TUM Kitchen Data Set of Everyday Manipulation Activities for Motion Tracking and Action Recognition. In *IEEE Int. Workshop on Tracking Humans for the Evaluation of their Motion in Image Sequences (THEMIS). In conjunction with ICCV2009*, 2009.

[17] Moritz Tenorth and Michael Beetz. KnowRob — Knowledge Processing for Autonomous Personal Robots. In *IEEE/RSJ International Conference on Intelligent RObots and Systems.*, pages 4261–4266, 2009.

[18] Moritz Tenorth, Lars Kunze, Dominik Jain, and Michael Beetz. KNOWROB-MAP – Knowledge-Linked Semantic Object Maps. In *Proceedings of 2010 IEEE-RAS International Conference on Humanoid Robots*, Nashville, TN, USA, December 6-8 2010.

[19] Ian H. Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2nd edition, 2005.

[20] Li Yujian and Liu Bo. A normalized levenshtein distance metric. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(6):1091 –1095, june 2007.

[21] H. Zender, O. Martínez Mozos, P. Jensfelt, G.J.M. Kruijff, and W. Burgard. Conceptual spatial representations for indoor mobile robots. *Robotics and Autonomous Systems*, 2008.