Hierarchical Segmentation of Manipulation Actions based on Object Relations and Motion Characteristics

Mirko Wächter and Tamim Asfour[1]

***Abstract*—Understanding human actions is an indispensable capability of humanoid robots which acquire task knowledge from human demonstration. Segmentation of such continuous demonstrations into meaningful segments reduces the complexity of understanding an observed task. In this paper, we propose a two-level hierarchical action segmentation approach which considers semantics of an action in addition to human motion characteristics. On the first level, a semantic segmentation is performed based on contact relations between human end-effectors, the scene, and between objects in the scene. On the second level, the semantic segments are further sub-divided based on a novel heuristic that incorporates the motion characteristics into the segmentation procedure. As input for the segmentation, we present an observation method for tracking the human as well as the objects and the environment. 6D pose trajectories of the human's hands and all objects are extracted in a precise and robust manner from data of a marker-based tracking system. We evaluated and compared our approach with a manual reference segmentation and well-known segmentation algorithms based on PCA and zero-velocity-crossings using 13 human demonstrations of daily activities. We show that significantly smaller segmentation errors are achieved with our approach while providing the necessary granularity for representing human demonstrations.**

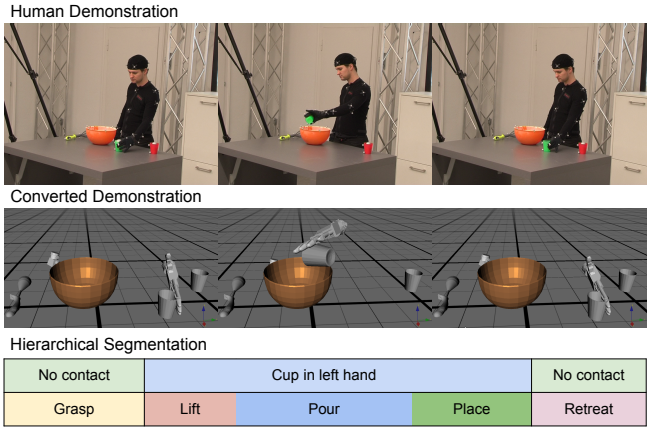

Fig. 1. A human demonstration of a complex task (top) is being recorded with a marker-based motion capture system. These marker-trajectories are converted into 6D object pose trajectories (middle), which serve as input for the proposed segmentation algorithms. The result of the segmentation (bottom) contains segments with distinct object-relations on the top-level and sub-segments with distinct motion characteristics on the bottom-level. The subsegments in this figure are labelled manually to illustrate the meaning of the subsegments.

I. Introduction

Research efforts in humanoid robotics have been dedicated to the development of sophisticated systems that can mimic the functionalities of a human. To tap this huge potential, humanoid robots are to be endowed with cognitive abilities for the acquisition of novel motor knowledge and the adaptation of this knowledge to unseen situations in order to account for dynamic changes. An intuitive way to approach this challenge is to acquire motor knowledge through the observation of humans and to transfer this knowledge to robots. In this context, an emerging paradigm is programming by demonstration [1], which in recent years progressed to the more biological-oriented term of imitation learning. A central concept which provides the basis for numerous imitation learning approaches has been the concept of the motion primitives. Motion primitives are units which incorporate a control policy for the execution of simple, basic motion patterns. Commonly, it is assumed that these motion primitives form the human motion repertoire from which complex movements are generated, adapting and sequencing these primitives in a task-dependent way. To provide data from which a robot can learn these motion primitives, methodologies have to be developed to allow the automatic segmentation of continuous human motion.

In particular, a segmentation procedure that provides a sequence of reliable segmentation points is needed. These segmentation points should denote changes in the scene which are caused by the enclosed manipulation actions. However, such points are difficult to extract from mere human motion data. Therefore, we consider in this work also motion data of the manipulated objects. In addition, we also wish to determine smaller segments depending on the elements the manipulation action is composed of.

To develop such a segmentation method which satisfies the demands mentioned aboves, in this work we present a hierarchical segmentation approach. Additionally, it should serve as an automatic tool to enrich a motion library with semantic information and more granular motion element. On the higher level a semantic segmentation is performed based on the contact relations between the human end-effectors, the scene and between objects in the scene. To enable the capturing of these relations, a method has been developed which allows the robust and accurate acquisition of human and object motion. On the lower level, the semantic segments are further analyzed in order to identify motion primitives. The proposed segmentation approach constitutes a crucial component in a motion learning framework.

The paper is organized as follows: Section II provides an overview of related work. In Section III, the acquisition of human motion data featuring demonstrations of complex manipulation tasks is described. In Section IV, the proposed action segmentation method procedure is introduced. The

[1]Mirko Wächter and Tamim Asfour are with Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Germany {waechter,asfour}@kit.edu

*The research leading to these results has received funding from the European Union Seventh Framework Programme under grant agreement Nº 270273 (Xperience).

approach is evaluated in Section V. Section VI summarizes the work and discusses future extensions.

II. Related Work

In order to be able to understand and analyse complex human motor behaviors, demonstrations of these behaviors have to be decomposed into meaningful segments which denote manipulation actions as well as the corresponding action primitives. For this purpose, in the field of robotics, a large number of different approaches have been proposed mainly in the context of imitation learning. In general, segmentation algorithms can be categorized into unsupervised and supervised methods. Unsupervised methods do not require any prior knowledge of the actions which are featured in the behavior to be segmented, and, thus, a temporal segmentation of continuous human movements can be performed in an online manner. According to this methodology, in [2], an approach has been introduced based on the joint velocities. Segments are enclosed between points where the mean squared velocity falls below a predefined threshold. In [3], this method has been extended in a way that this critical threshold is determined based on the scaling of the current mean squared velocity. In addition, tactile feedback has been incorporated in order to detect changes in the contact relations between human and environment while the human is in motion. In [4], a method based on zero velocity crossings is proposed. Using this method, segments are denoted by points where in a sufficient number of joints the movement direction is changing. A probabilistic approach has been proposed by [5] using Principal Component Analysis in order to identify segments represented in a low-dimensional space. The reconstruction error for newly observed movements indicates whether the observation belongs to a current segment or denotes the beginning of a novel one. In [6] a framework for learning actions from observation is proposed, in which also the agent and the objects are tracked and state changes considered. An interaction learning system is proposed in [7], which uses the relative velocity between an object and the hand as the segmentation feature. However, both works do not consider motion characteristics for segmentation.

In contrast to unsupervised approaches, the segmentation with supervised methods is based on previously known segments mostly represented in a generalized form. Regarding the learning of human actions, a common strategy is to use the same representation for the learning as well as segmentation and recognition of actions and motor primitives.

In [8], [9] and [10], segments are identified and represented as states of an Hidden Markov Model (HMM). In [8], the segmentation points are derived by optimizing the cost path using a modified Viterbi algorithm. To gradually verify and to refine the segmentation, prior knowledge is introduced in the form of HMMs representing known segmented movements which are grouped by applying a clustering procedure. In [9] the authors propose a Mimesis-model that abstracts whole body motion patterns as symbols and allows segmentation and recognition of motions. Though, they state it is only applicable to simple motions. An alternative approach using dynamical systems is proposed in [11]. Linear time invariant dynamical systems are designed and trained to represent specific drawing primitives. For novel observations, parameters for these systems are estimated. Based on the parametric

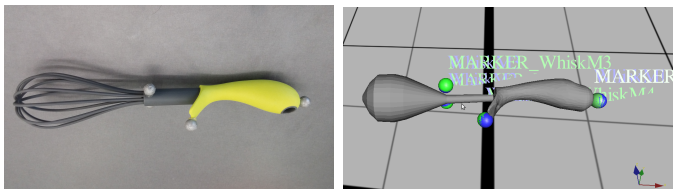

Fig. 2. Left: Image of a real object with attached reflective markers. Right: A visualization of the corresponding 3D mesh model with attached virtual markers (blue/green spheres).

error and the corresponding approximation error segments are identified. In [12], latent force models are used in order to segment human movements. Multiple dynamical systems are used to encode the relations between a latent force space and the joint movements. Smooth trajectories in the latent force space indicate possible segments. However, the segmentation result strongly depends on the dimensionality of the force space. A further approach which uses the Dynamic Movement Primitives (DMP) for segmentation and recognition of basic movements is introduced in [13]. Based on a library of previously trained DMPs representing various basic actions, the observation of an action is encoded as a novel DMP. If this DMP matches an element in the library a segment is found. Otherwise, the novel DMP is used to update the library.

Compared to unsupervised methods, supervised approaches yield more accurate results for movements which are already known to the system. To ensure the robustness of these approaches the prior knowledge should consider segments which can be sufficiently discriminated, and, thus, found segments incorporate a certain complexity. Therefore, unsupervised methods are better suited for a fine-granular decomposition of a manipulation action in motion primitives, especially if the primitives are new to the system. However, finding the segmentation parameters such as thresholds which yield an optimal result is difficult. To our best knowledge, none of the approaches above provide a solution for motion segmentation which takes into account both the semantic of the task as well as motion characteristics of the demonstration. With our hierarchical segmentation approach we pursue novel segmentation strategies which consider both semantic and motion information of human demonstration (see Fig. 1).

III. Action Data Acquisition

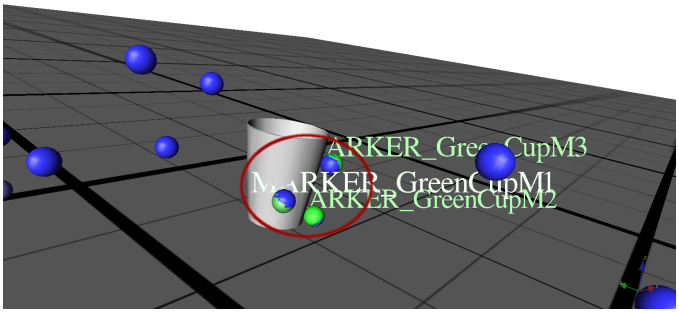

Fig. 3. Mapping of observed markers (blue spheres) attached to the cup to virtual markers on the model. Both observed and virtual markers are aligned (green/blue spheres in red circle). The visualization of the cup's 3D mesh shows the 6D pose of the perceived object.

In this section, we will address how action data can be captured in a reliable way. In order to capture human demonstrations a variety of different manipulation actions with a high accuracy and at a high resolution, we employ a marker-based motion capture system [14]. The demonstrations involve not only a human, but also the objects, which are manipulated by the human.

To be able to capture both, markers have been attached to the human subject as well as to objects of interest located in the current scene (see Fig. 2). For simplicity reasons, only the hands of the human are considered during the capturing process and are treated as rigid bodies similar to the objects. On each object and hand at least three markers are attached, although more markers increase the robustness in case of occlusions. All markers are labeled and grouped according to the object they belong to. The motion capture data contains Cartesian space trajectories of all markers. In our previous work [15], we used only the trajectories of marker groups for the segmentation. However, the arrangement of the markers do not sufficiently describe the shape of the object, and, thus, the pose of the object can not be inferred based on the mere marker positions. To deal with these shortcomings, we convert the trajectories from a marker representation to a 6D pose representation for each object with the help of the Master Motor Map framework [16]. The first component of the representation are 3D mesh models. Thus, we create 3D mesh models of each object with a 3D object scanner [17] or a common 3D modeling tool and attach virtual markers on the model (see Fig. 2). The resulting object representation consists of 3D mesh model of each object as well as the positions of all markers attached to this object. This information allow to calculate the 6D object pose. For this point set registration problem we applied the approach presented by Besl in [18]. In Fig. 3 the mapping of observed and virtual markers is visualized.

To retrieve the 6D pose trajectory for an object, the transformation is calculated for each frame and applied to the base pose of the object, which is usually the identity. This conversion is done for all captured objects, which leads to an accurate representation of the scene during the demonstration. In our applications, we use trajectories with at least three markers for each object in any frame.

IV. Hierarchical Action Segmentation

We propose in this paper a two-level hierarchical segmentation method for segmenting a complex task into meaningful segments. On the top level, segments are identified based on a semantic criteria. On the bottom level, segments are identified on a heuristic for motion characteristic. Semantic segments are identified based on contact relations between the human end-effectors and objects in the scene. This top level of segmentation is performed based on the captured 6D trajectories of objects and end-effectors as well as on the 3D object models. On the bottom level, the resulting semantic segments are further analyzed regarding their motion characteristics to identify a sequence of motion primitives.

A. Semantic Segmentation based on Object Contact Relations

First, the human demonstration is segmented based on the 6D pose trajectories of each object while making use of the 3D model information of each object. The segmentation is grounded on the spatial relations between the objects. A similar approach was proposed by Aksoy et al., which uses RGB stereo camera [19], [20] or RGB-D [21] images as input. While this approach is model-free, it does not provide 6D trajectories of human end-effectors or objects. The authors estimate contacts between objects by recognizing overlapping color blobs. In our previous work in [15], we utilized only marker distances for contact detection in demonstrations. As mentioned in Section III, the shape and the pose of the object are not sufficiently represented by the markers alone. A segmentation based on the distances between the objects requires the use of high distance thresholds to detect contact points that are far from the markers. This reduces the robustness of such a method since objects in the demonstrations need to have a relatively large minimum distance between each other. The introduction of a 3D mesh model as object representation instead of markers allows the use of sophisticated mesh-based collision detection algorithms, such as the one described in [22], to accurately calculate the distance between objects and to detect contacts/collisions between them.

The demonstration is segmented by detecting key frames which are extracted based on the change of relations between objects and the human hands, which correspond to the end-effectors in our manipulation tasks. We consider only contact relations, where $contact(A, B)$ denotes contact between object A and B. Other relations like on or in ground on such contact relations and are not relevant for our segmentation method. For future extension towards symbolic planning, the incorporation of further relations might be useful. The relation $contact(A, B)$ relies on the closest distance between any part of the involved objects A and B. For each frame of the demonstration, the relations between all objects are calculated and key frames are stored whenever the relation between two object changes its status. $contact(A, B)$ returns true if the distance falls below a predefined threshold. To deal with noise on the distance measure, we use hysteresis on the threshold, which is increased when a contact has been detected in the last frame. This results in a sequence of key frames. Additionally, the world state is stored together with every key frame. This world state is the set of all relations between all objects. It describes the current status of the scene and can be used for association with known actions (as described in [15]). As stated before, every relation change leads to a key frame. However, not all actions correspond to only one relation change. For example, the action of *pouring* liquid (L) in a cup (C) into a bowl (B) can be associated with two relation changes (under the assumption that liquid can be tracked):

$$\begin{array}{l} contact(L,C) \quad \wedge \\ !contact(L,B) \end{array} \quad \rightarrow \quad \begin{array}{l} !contact(L,C) \quad \wedge \\ contact(L,B) \end{array}$$

where L stands for liquid, C for cup and B for bowl. Hence, key frames need to be merged into groups of key frames that semantically belong together. For most actions, these key frames appear with a small delay between each other as it is the case e.g. for the action of *dropping* an object into another. A simple way to cope with this is using the temporal displacement of two key frames to merge them once this distance is too small. State changes are always instantaneous, although e.g. *pouring* might seem to take time. However, the change of the contact relation does not take time. If the *pouring* would take noticeably long, it would result in two

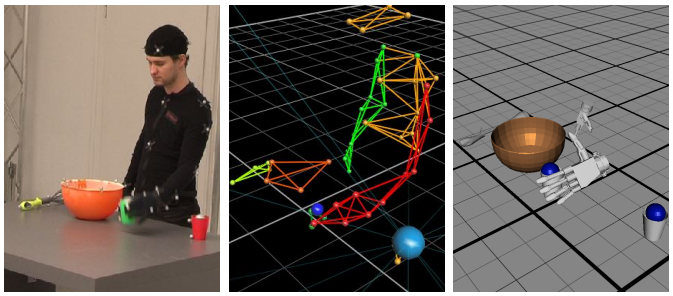

Fig. 4. Stages of demonstration capturing and processing: The human demonstrator with attached markers on the objects (left); marker group representation (middle); 3D mesh models with applied 6D object pose (right).

key frames: a first key frame when the liquid gets in contact with the target container and a second key frame when the liquid loses contact with the source container.

B. Segmentation based on Motion Characteristic

This first level of segmentation described above results in a segmentation of the human demonstrations in semantic segments, that have observable changes in the world state. However, some actions have unobservable effects, even for a human. For example, the effect of shaking two transparent liquids in a bottle cannot be observed visually. In this and other examples, such unobservable effects are relevant for segmentation and finally understanding the demonstrated task. The number of these unobservable effects for many tasks can exceed the number of observable effects. However their detection based on current state of the art methods and sensor technologies is challenging.

The previously described method can only detect moments when two objects make contact with each other. The aforementioned effect and therefore the state change cannot be detected. To this end, we extend the segmentation from the previous section with a subsegmentation that extracts motion parts within a semantic segment based on the trajectory shape and the motion characteristics. In other words, we take the detected semantic segments as the input for the bottom-level and divide them further. The goal of the subsegmentation is to split up the semantic segments into smaller parts that contain motions with different motion characteristics and potentially represent the different motion primitive within a semantic segment. Several motion-based segmentation methods could be used for this subsegmentation. In this work, we use a heuristic that incorporates the characteristic of a motion into the approach. To capture the characteristic of a motion, our approach uses as a basis the dynamics of the motion, i.e. the acceleration values of the trajectory.

There are two fundamentally different ways to segment motion data. One is to find key frames that meet a specific criteria, and the other one is to search for meaningful segments. Our approach lies in between. The approach searches for key frames that maximize the difference of the trajectory parts left and right of this key frame. As such, the approach differs from the pure key frame search since the key frame itself is unimportant. It also differs from segment search because it does not require the complete demonstration segments to be known in advance. In short, our approach segments the trajectory in most distinctive parts. To find the key frames, the demonstration trajectory of is analyzed recursively. On every recursion level, the given trajectory segment is searched sequentially with a predefined step size for the key frame, that divides the trajectory best. Subsequently, the segments left and right of this key frame candidate are analyzed again in the same manner until the segment size falls below a threshold or no additional segments with a sufficiently good quality can be found. The whole approach is described in Algorithm 1.

To define the quality of a frame, which is needed to decide whether a frame is a key frame, we introduce first the following terms:

$$A_d(t) = |a_d(t+1) - a_d(t)| \tag{1}$$

$$s_{l,d}(t_c) = \sum_{t=t_c-\frac{w}{2}}^{t_c-1} A_d(t) \left(\frac{\hat{U}_l}{\hat{U}_r}\right)^2 \tag{2}$$

$$s_{r,d}(t_c) = \sum_{t=t_c}^{t_c+\frac{w}{2}-1} A_d(t) \left(\frac{\hat{U}_r}{\hat{U}_l}\right)^2, \tag{3}$$

where $a_d(t)$ is the acceleration vector of dimension d at timestamp t, d is the dimension of the trajectory, $s_{l,d}(t_c)$ is the score left of the key frame candidate, t_c is the timestamp of the key frame candidate, w is the window size left and right of the key frame candidate that is analyzed, $\hat{U}_l$ and $\hat{U}_r$ is the peak-to-peak amplitude respectively left and right of the key frame candidate. Eq. (2) calculates the score of the segment left of the key frame by calculating basically the length of the function. Eq. (3) does the same for the right side of the key frame. To also consider the amplitude of the acceleration, the score is multiplied with the squared relation of the peak-to-peak distances left and right of the key frame candidates. Finally, the quality q_d of a key frame candidate is then defined as:

$$q_d = \begin{cases} s_{l,d}/s_{r,d} & s_{l,d} > s_{r,d} \\ s_{r,d}/s_{l,d} & s_{l,d} \leq s_{r,d} \end{cases}. \tag{4}$$

Until now the qualities for each dimension are normalized to their amplitudes. However, the amplitude of one dimension can be small compared to another dimension. Since motions in a dimension with overall low amplitudes are not as important as another dimension with high amplitudes, the qualities for each dimension are aligned with the maximal peak-to-peak distance $\hat{U}_d$ of all dimensions:

$$\hat{q}_d = q_d \cdot \sqrt[z]{\frac{\hat{U}_d}{\max_d \hat{U}_d}}, \tag{5}$$

where z is a scalar to influence the weight of the normalization. The best $\hat{q}_d$ of all frames and dimensions is selected as a key frame with the quality q, if the value does not violate a quality-threshold or a minimum segment size to avoid oversegmentation. The idea for this heuristic is that motions with a different characteristic, e.g. smooth circles, intense shaking, pouring, etc. have a different acceleration profile and therefore a different shape. The heuristic primarily measures the length of the acceleration curve and normalizes it with the amplitude of the segment.

Algorithm 1 Motion Characteristic Segmentation Algorithm

```
function FINDKEYFRAMES(kf, t_l, t_r, l_min)
    // kf: in-out parameter; intially empty key frame list
    // t_l, t_r: timestamps of current segment borders
    // l_min: minimum segment length
    for t := t_l + l_min to t_r − l_min ; t += 0.01 do
        for d := 0 to dimensions do
            q_n ← CALCQUALITY(t, d)
            if q_best < q_n then
                q_best ← q_n
                t_best ← t
            end if
        end for
    end for
    if q_best > λ then
        kf.INSERT(t_best,q_best)
        FINDKEYFRAMES(kf,t_l,t_best,l_min)
        FINDKEYFRAMES(kf,t_best,t_r,l_min)
    end if
end function
```

V. Experiments and Evaluation

In this section, the experimental setup for data acquisition is described and the results of conversion of marker positions into 6D object poses is discussed. We compare the segmentation results of our approach with other segmentation methods based on a new segmentation metric. Using this metric we compare our approach to manual segmentation and other segmentation algorithms. We describe the experiments we conducted and results we achieved by using the hierarchical segmentation based on object contact relations and the subsegmentation based on shape and motion characteristics.

A. Experimental Setup

The marker-based motion capture system consists of 10 cameras for the observation of the scene, in which all objects are common rigid household objects that have at least three markers attached to them in an asymmetric arrangement. To capture human motion during the demonstration, only markers at the hand were considered where the hands were treated as rigid bodies as well. The motion capture system contains models of the spatial marker relations of all objects allowing the automatic labeling of the markers (see Fig. 4, middle). For every object, a 3D mesh model was either created with a 3D scanner or by hand and extended with virtual marker positions (see Fig. 2).

B. Segmentation Metric

To compare our approach to the other algorithms, we propose a metric that measures the error of the segmentation in square seconds compared to a reference segmentation. Let K_r be the set of key frame of the reference segmentation and K_f the found key frames of the algorithms. The metric assigns for each $k_r \in K_r$ the closest key frame available in K_f. Each key frame can only be assigned once and measures the squared error to the reference key frame. The maximum allowed distance for a correct key frame to the reference key frame was chosen to be 1 second, otherwise the key frame is considered *missed*. For every missed key frame and for each false positive key frame of the algorithms, a penalty $p = 7\ s^2$ is added, so that completely wrong key frames are severely penalized. In summary, the metric we use for comparison is given as:

$$e = (m + f) \cdot p + \sum_i \min_j (k_{r,i} - k_{f,j})^2, \tag{6}$$

where m is the number of missed key frames and f the number of false positives.

C. Experiments

To test our approach we recorded the following demonstration scenarios of action sequences: *preparing batter*, *wiping a table*, *shaking and pouring a bottle's content into a bowl* and *polishing a bowl*. The *Preparing batter* scenario contains two cups, which are grasped by the human, who pours the content of both cups into a bowl and then places the cups again on the table. Afterwards the liquids are mixed with a whisk, which also has to be grasped by the human to perform the mixing task, and which is eventually placed on the table. This scenario was chosen because it contains several objects and typical actions in the context of household robot. In Fig. 5 the semantic segmentation of one trial of this scenario is depicted. The inter-object distances that do not matter for this segmentation task are omitted in this diagram for better clarity. Whenever the left or right hand grasps an object or puts down an object a new key frame is inserted. In the *table wiping* scenario, the human demonstrator grasps a sponge from a table and wipes the table using several different wiping styles like intensive wiping of a spot or wiping of a big area with circles. In the third scenario, a bottle is being grasped, tossed, inspected, shaken, poured and dripped off. In the fourth scenario, a big bowl is being held in one hand and polished by the other hand with changes of the hands in between. In each scenario the trials vary in the selection, duration and order of the actions.

D. Evaluation and Comparison

We evaluated our approach on a set of 13 action sequences in the previously described scenarios. Generating ground truth data from these demonstrations is difficult since the actions transition often smoothly into the next action without a clear cut. Also, it is even for the human not clear when precisely an action ends and starts. Nonetheless, the recordings of these action sequences were segmented manually as a reference to which the results of the algorithms are compared. Results of the pure motion characteristic segmentation are demonstrated in Fig. 6 on a motion that changes the motion characteristics frequently.

In Fig. 7 the results of the hierarchical segmentation for the shaking-pouring scenario in comparison with manual reference segmentation, Principal Component Analysis (PCA) and Zero-Velocity-Crossings (ZVC) are shown. The two types of input values (position and object relations) are both depicted in bottom of the figure. The distance between objects serves as input for the semantic segmentation and the position values are the input for the other methods. In the middle of the figure the found key frames for each algorithm are denoted. For the hierarchical segmentation, the key frames found in the separate levels are visualized with a different color. In this example, it can be seen how important the segmentation based on the

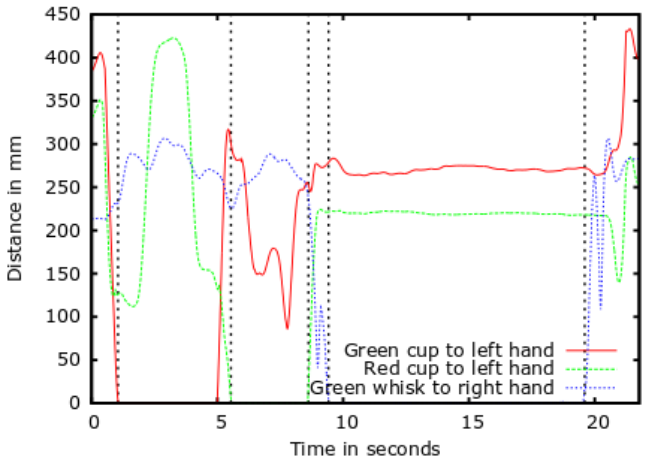

Fig. 5. Segmentation based on object contact relations: When two objects get in contact with each other (contact in this case is approximated as $distance < 7mm$) or lose contact, a new key frame is inserted with the current world state attached to it. The dotted vertical lines depict the detected semantic segments. Only distances between objects that get in contact during the complete demonstration are shown.

motion characteristic is. The semantic segmentation only found two key frames (green bars), because object relation changes only occurred twice: when grasping and placing the bottle. The motion based segmentation found eleven key frames, from which 7 are also found in the manual segmentation. This shows, depending on the action sequence, that the hierarchical segmentation significantly improves the pure semantic segmentation from our previous approach [15].

The average metric results over all 13 action sequences are shown in Table I. In this table, our Hierarchical Segmentation (HS) algorithm is compared using the proposed metric with segmentation algorithms based on Zero-Velocity-Crossing (ZVC) and Principal Component Analysis (PCA). In addition, all 13 demonstrated action sequences (30-40 seconds each) are

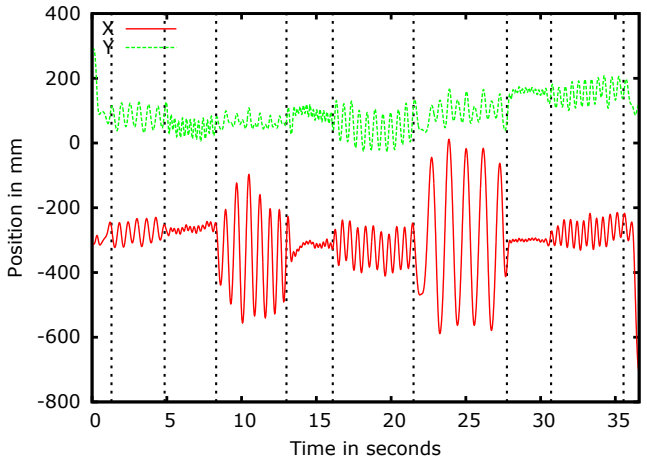

Fig. 6. Subsegmentation of one semantic segment (sponge touches hand and table) into different subsegments, i.e. different wiping styles. The z-dimension is omitted in the figure. The dotted vertical lines depict the segmentation points. Whenever the wiping styles changes, a key frame is inserted.

Average Results	HS	ZVC	PCA
Error	3.35 s^2	7.01 s^2	20.18 s^2
Accuracy	0.27 s	0.1 s	0.36 s
Unmatched key frames	2	0.6	27.9
Missed key frames	3.54	12.27	3.5

TABLE I. COMPARISON WITH OTHER SEGMENTATION METHODS.

compared to manual segmentation performed by two persons. The error row indicated the value of the proposed metric. Accuracy is the average distance of matched key frames to the manual segmented key frames. Unmatched key frames (false positives) mean how many key frames found by the algorithms were not assigned to a key frame denoted in the manual segmentation. Missed key frames (false negatives) indicate how many key frames of the manual segmentation where not assigned to a key frame found by the algorithms. The results are the average results over the 13 action sequence demonstrations. The parameters for our approach were trained on five action sequences with a genetic algorithm and tested on all 13 sequences.

It can be seen that the proposed algorithm achieves significantly smaller error values than the two other approaches. In Fig. 8 three trials of the shaking-pouring scenario are shown in comparison to the manual segmentation and the hierarchical segmentation. Most of the key frames are found and only a few are missed.

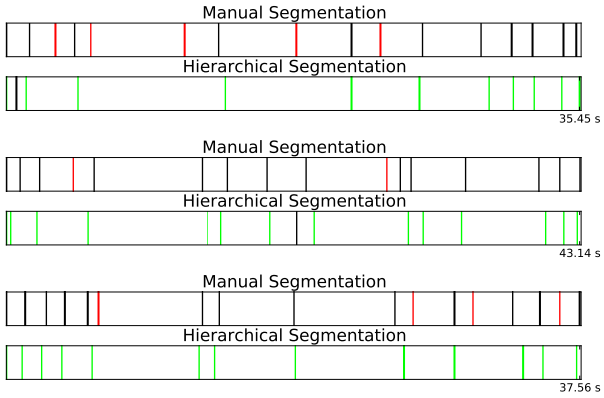

Fig. 8. Comparison of manual to hierarchical segmentation with 3 trials of the same shaking-pouring scenario. The vertical lines denote key frames at their timestamps. Missed key frames are shown in red. Matched key frames are shown in green. The 3 trials contain grasping, placing, shaking, tossing, inspecting and pouring actions similar to Fig. 7 (top) each in different order and with different timing.

The difficulties in these scenarios lie in the unobservable effects of some of the actions, e.g. in *pouring* and *mixing* since the liquids are not tracked. Therefore, the actions cannot be detected with the semantic segmentation based on object-relation changes and a detection on motion characteristic level is required.

E. Discussion

In the following we discuss our achievements and results from different relevant point of views.

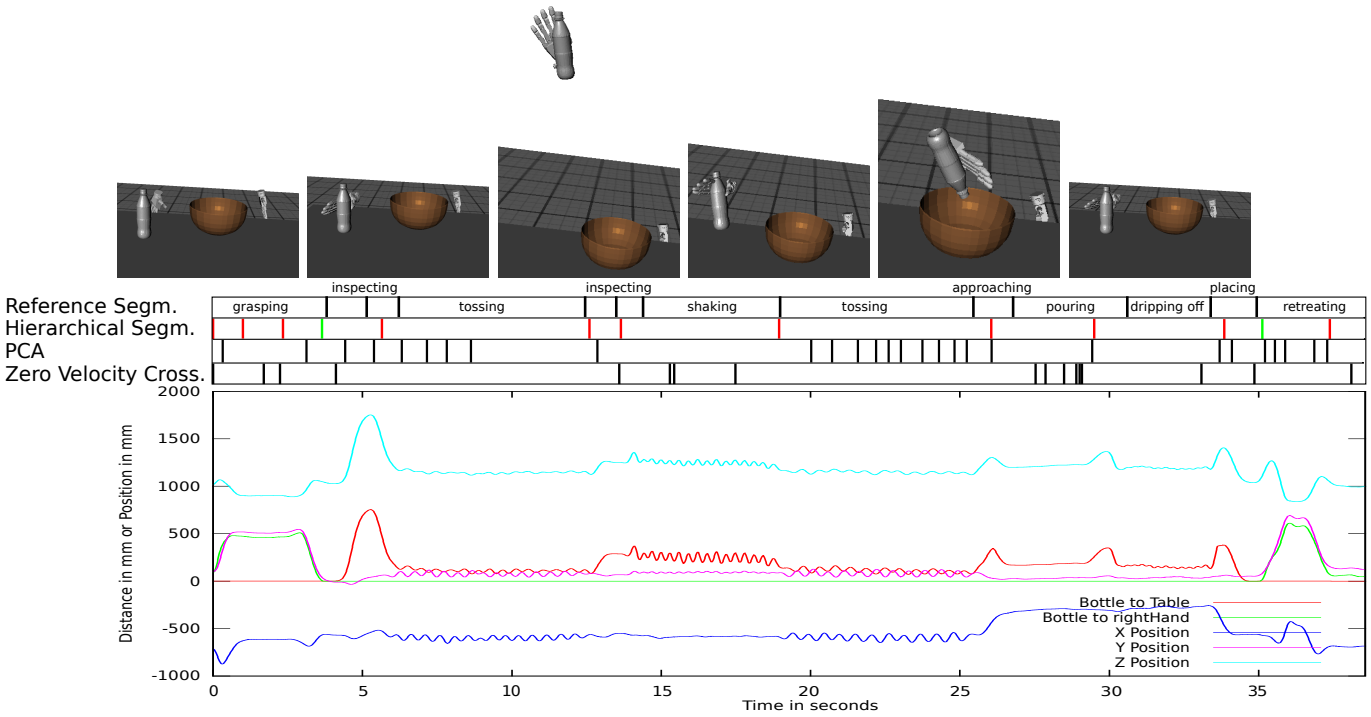

Fig. 7. Segmentation of an action sequence with multiple algorithms compared to a manual reference segmentation: Top: Snapshots of the action sequence in chronological order of the analyzed sequence for visualization. Middle: Comparison of a manual reference segmentation, PCA, Zero-Velocity-Crossing and the proposed hierarchical segmentation of an action sequence shown with the key frames (vertical bars). The red bars in the Hierarchical Segmentation denote the key frames found by the motion characteristic heuristic and the green bars denote the semantic key frames. The action sequence contains grasping, placing, shaking, tossing, pouring, inspecting, dripping off of a bottle and placing. Bottom: Position of the right hand in Cartesian coordinates and relevant distances between objects during the task execution.

1) Data acquisition: Evaluating the precision of the object tracking is difficult since there is no ground truth data. However, in general, the algorithm will produce exact solutions unless the input data is contaminated with noise or error. The precision depends on the following components:

- Positional precision of the markers. With the used motion capture system, the deviation lies around 1 mm.
- The position of the virtual markers. This is in practice a common source of error, since the marker are placed by hand on the object in a 3D modeling tool.
- The 3D model of the object. Errors in the modeling the object affect the contact detection of the objects.
- Markers on non-rigid bodies like hands do not have constant relative positions to the other markers on the object, if the object is transformed and therefore impede the calculation of the 6D pose. The marker registration algorithm minimizes the error if there are more than three markers. However, an error will persist. In our application, we treat the hands as rigid bodies for simplicity, as the error is small enough to deal with.

2) Semantic segmentation: Compared to our previous work in [15], the segmentation results showed significant improvements. The segmentation in [15] has been done solely based on marker positions, which do not represent the object shape accurately and therefore contact detection was not reliable. A high threshold was needed for an approximated contact detection, which easily lead to false-positives. This deficiency has been alleviated by incorporating the region around every key frame. Since our new approach makes use of 3D object models, the contact detection threshold can be reduced to a $\frac{1}{20}$th of the old threshold (7 mm instead of 150 mm). In general, the new segmentation approach relies strongly on the precision of the 6D trajectory and the model associated with it. An inaccurate model can lead to false or missed contact detection and therefore to false segmentation. Based on our experience, the precision is sufficient to detect all contacts and segment the trajectory into their semantic parts. As shown in Fig. 5, where the action sequence of pouring with two different cups and a subsequent mixing action was demonstrated, the contact between the objects can be extracted from the distance-curve and, thus, the key frames are inserted. For the sake of clarity, several distance curves between the objects have been omitted. In certain cases, a lost contact does not mean that a new action starts. For example, during the wiping action, the sponge occasionally loses contact with the table. Based on the assumption that actions have minimum duration (we set 500 ms as a threshold), this situation is avoided to a certain degree by merging adjacent key frames that are temporally too close to each other.

3) Subsegmentation based on Motion Characteristic: The subsegmentation tackles the problem of unobservable effects of actions. Additionally, different styles of periodic actions

can be detected (different wiping styles, e.g. wiping in lines or intensive wiping on one spot). In Fig. 6, a segmented action of wiping is shown, which is then subsegmented in different wiping styles. Each time the wiping pattern changes, a new key frame is inserted. In Fig. 7, the further inspection and segmentation of a semantic segment, which supposedly represents a pouring action, indicates these segments comprise more actions without observable effects, and, thus, can be further divided into subsegments. In our experiments, most segments have been detected, though a smooth transition between two actions can be problematic and no key frame might be found.

VI. Conclusion

In this work, a hierarchical segmentation approach has been presented which first allows the reliable determination of semantic key points in continuous human movements. These points denote clear transitions between different manipulation actions where the human changes the scene. To detect these changes, a robust method has been implemented which enables the robust and accurate pose estimation of objects and the human hands. This is then used for the detection of contact relations between these entities based on mere marker-based motion capture data. We showed that such semantic segmentation cannot deal with unobservable effects of actions when important key frames of the demonstration can neither be determined algorithmically nor detected by current sensor technologies. To this end, we propose a new algorithm for the subsegmenting of semantic segments based on trajectory shape and motion characteristics. The algorithm is based on a heuristic that incorporates the characteristic of a motion into the segmentation procedure. The algorithm differs from other segmentation approaches in the literature as it represent a compromises between key frame search based on certain criteria, but does not require the complete trajectory to be known in advance.

The experimental evaluation shows that the hierarchical segmentation approach allows the identification of meaningful segments in complex human demonstrations without over-segmentation and without omitting important demonstration key frames. Due to the hierarchical approach, the found segmentation points are enriched with additional information which can be useful for the organization, the sequencing, and the reproduction of learned actions and motion primitives. Future work is concerned with the extension of the underlying heuristic for motion segmentation. Further, we will investigate the use of the segmented demonstrations as input for our work on imitation learning of manipulation action primitives.

References

[1] A. Billard, S. Calinon, R. Dillmann, and S. Schaal, *Survey: Robot Programming by Demonstration*, 2008.

[2] M. Pomplun and M. J. Matarić, "Evaluation metrics and results of human arm movement imitation," in *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*, 2000.

[3] J. Lieberman and C. Breazeal, "Improvements on action parsing and action interpolation for learning through demonstration," in *4th IEEE/RAS International Conference on Humanoid Robots*, vol. 1. IEEE, 2004, pp. 342–365.

[4] A. Fod, M. J. Matarić, and O. C. Jenkins, "Automated derivation of primitives for movement classification," *Autonomous robots*, vol. 12, no. 1, pp. 39–54, 2002.

[5] J. Barbic, A. Safonova, J. Y. Pan, C. Faloutsos, J. K. Hodgins, and N. S. Pollard, "Segmenting motion capture data into distinct behaviors," in *GI '04: Proceedings of Graphics Interface 2004*. School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada: Canadian Human-Computer Communications Society, 2004, pp. 185–194.

[6] V. Krüger, D. Herzog, S. Baby, A. Ude, and D. Kragic, "Learning actions from observations," *Robotics & Automation Magazine, IEEE*, vol. 17, no. 2, pp. 30–43, 2010.

[7] M. Mühlig, M. Gienger, and J. J. Steil, "Human-robot interaction for learning and adaptation of object movements," in *IEEE/RSJ International Conference on Intelligent Robots and Systems, October 18-22, 2010, Taipei, Taiwan*, pp. 4901–4907.

[8] D. Kulic and Y. Nakamura, "Scaffolding on-line segmentation of fully body human motion patterns," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2008, pp. 2860–2866.

[9] T. Inamura, I. Toshima, H. Tanie, and Y. Nakamura, "Embodied symbol emergence based on mimesis theory," *The International Journal of Robotics Research*, vol. 23, no. 4-5, pp. 363–377, 2004.

[10] T. Asfour, P. Azad, F. Gyarfas, and R. Dillmann, "Imitation learning of dual-arm manipulation tasks in humanoid robots," *International Journal of Humanoid Robotics*, vol. 5, no. 2, pp. 183–202, December 2008.

[11] D. Del Vecchio, R. M. Murray, and P. Perona, "Decomposition of human motion into dynamics-based primitives with application to drawing tasks," *Automatica*, vol. 39, no. 12, pp. 2085–2098, Dec. 2003.

[12] M. Alvarez, J. R. Peters, N. D. Lawrence, and B. Schölkopf, "Switched latent force models for movement segmentation," in *Advances in neural information processing systems*, 2010, pp. 55–63.

[13] F. Meier, E. Theodorou, and S. Schaal, "Movement segmentation and recognition for imitation learning," in *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

[14] J. Lee, J. Chai, P. S. Reitsma, J. K. Hodgins, and N. S. Pollard, "Interactive control of avatars animated with human motion data," in *ACM Transactions on Graphics (TOG)*, vol. 21, no. 3. ACM, 2002, pp. 491–500.

[15] M. Wächter, S. Schulz, T. Asfour, E. Aksoy, F. Wörgötter, and R. Dillmann, "Action sequence reproduction based on automatic segmentation and object-action complexes," in *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*, 2013, pp. 189–195.

[16] O. Terlemez, S. Ulbrich, C. Mandery, M. Do, N. Vahrenkamp, and T. Asfour, "Master Motor Map (MMM) - Framework and Toolkit for Capturing, Representing, and Reproducing Human Motion on Humanoid Robots," in *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*, 2014.

[17] A. Kasper, Z. Xue, and R. Dillmann, "The kit object models web database: An object model database for object recognition, localization and manipulation in service robotics," in *The International Journal of Robotics Research*, 2012.

[18] P. J. Besl and N. D. McKay, "A method for registration of 3-d shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 239–256, Feb. 1992.

[19] E. E. Aksoy, A. Abramov, F. Wörgötter, and B. Dellen, "Categorizing object-action relations from semantic scene graphs," in *ICRA*, 2010, pp. 398–405.

[20] E. E. Aksoy, A. Abramov, J. Dörr, N. Kejun, B. Dellen, and F. Wörgötter, "Learning the semantics of object-action relations by observation," *The International Journal of Robotics Research (IJRR)*, 2011.

[21] E. E. Aksoy, M. Tamosiunaite, and F. Wörgötter, "Model-free incremental learning of the semantics of manipulation actions," *Robotics and Autonomous Systems*, 2014.

[22] E. Larsen, S. Gottschalk, M. C. Lin, and D. Manocha, "Fast proximity queries with swept sphere volumes," Technical Report TR99-018, Department of Computer Science, University of North Carolina, Tech. Rep., 1999.