An Evaluation of Action Segmentation Algorithms on Bimanual Manipulation Datasets

Andre Meixner, Franziska Krebs, Noémie Jaquier, and Tamim Asfour

***Abstract*— Humans naturally execute many everyday manipulation actions with both arms simultaneously. Similarly, endowing robots with bimanual manipulation task models is key to efficiently perform complex manipulation tasks. To do so, a promising approach is to learn a library of task models from human demonstrations. However, this requires human motions to be meaningfully segmented. In this paper, we propose to segment the motion of each hand individually to account for different bimanual coordination patterns and provide a thorough evaluation of state-of-the-art segmentation algorithms on bimanual manipulation datasets. In particular, we compare segmentation algorithms at trajectory and semantic level with hierarchical algorithms. Moreover, our evaluation extensively studies the performances of various segmentation algorithms over a novel extension of the KIT Bimanual Manipulation Dataset featuring ∼176 minutes of human motion recordings in household scenarios.**

I. Introduction

Assistive robots providing help for humans in daily tasks should be able to learn new skills in an intuitive way and adapt them to new situations. Promising approaches to achieve these goals are Learning from Demonstrations (LfD) [1] and imitation learning. Humans execute many everyday actions with both arms simultaneously: Coordinating both arms according to different bimanual strategies [2] allows us to be more efficient and to accomplish more complex tasks. For example, in various household tasks such as cutting vegetables, the dominant hand performs the manipulation, while the non-dominant hand stabilizes the object being acted upon. Such tasks are significantly more difficult to perform with a single hand. In this sense, considering bimanual coordination and actions is key for robot manipulation [3]. However, learning bimanual manipulation task models from human demonstrations presents additional challenges and must be addressed at several levels, starting from approaches that enable segmentation of bimanual human demonstrations — both on semantic and on trajectory level.

Given a pre-existing skill library, the action segmentation problem can be reduced to a recognition problem [4]. However, creating a comprehensive library of skills or task models initially requires the motion to be segmented. Such segmentation can be achieved via (continuous) manual annotations [5], [6], via predefined heuristics [7]–[9], or via supervised [6], [10], weakly supervised [11] or unsupervised [12], [13] learning. Although these approaches have been extensively leveraged to segment unimanual and whole-body trajectories, comprehensive evaluations in the context of bimanual manipulation are still missing.

In this paper, we provide an extensive evaluation of state-of-the-art segmentation algorithms in bimanual manipulation scenarios (Section III). To do so, we propose to segment the motion of each hand *individually* to account for the fact that hands can execute uncoordinated actions, or coordinated actions spread along different time intervals [2]. We consider algorithms that segment motions *(i)* at trajectory level, *(ii)* at semantic level, i.e., considering contact changes between the human and objects, and *(iii)* by hierarchically combining trajectory and semantic levels. When required, we adapt these segmentation algorithms to consider bimanuality. As action segmentation via manual annotations requires significant manual efforts, and supervised learning approaches based on, e.g., graph networks [6], [10], do not generalize well to new environments or actions, we specifically focus on heuristics and unsupervised methods. We thoroughly evaluate and compare various segmentation algorithms on the human motions of the KIT Bimanual Manipulation Dataset [5], which contains 12 short bimanual household actions with a large number of variations (Section V). To evaluate the segmentation algorithms within more complex scenarios, we additionally extend this dataset to include 90 recordings of 3 long bimanual manipulation household tasks composed of sequences of the aforementioned manipulation actions (Section IV). Our evaluation shows that all segmentation algorithms generally display similar characteristics across short and long sequences of actions, and across subjects. Moreover, it confirms the benefits of hierarchical segmentation algorithms, while opening the door to further developments in segmenting bimanual actions and learning bimanual task models.

The contributions of this paper are twofold: *(i)* We benchmark various heuristics and unsupervised segmentation algorithms at trajectory and semantic levels, as well as hierarchical segmentation algorithms on bimanual manipulation actions and sequences thereof; and *(ii)* we extend our KIT bimanual manipulation dataset [5] with recordings of 90 annotated sequences of bimanual manipulation actions for six individuals. This extension is publicly available at `https://motion-database.humanoids.kit.edu/details/datasets/3521/`. A video of the recorded tasks and segmentation results accompanies the paper and is available at `https://youtu.be/VRccEiYhc-4`.

The research leading to these results has received funding from the European Union's Horizon Europe Research and Innovation programme under grant agreement No 101070292 (HARIA) and the German Federal Ministry of Education and Research (BMBF) under the project CATCH-HEMI (01KU2012). The authors are with the Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Germany {andre.meixner, franziska.krebs, noemie.jaquier, asfour}@kit.edu

II. Related Work

We first review segmentation approaches with a focus on bimanual manipulation. To do so, we categorize the different approaches into algorithms for *(i)* segmentation at *trajectory level*, *(ii)* segmentation at *semantic level*, and *(iii) hierarchic* segmentation which combines both. Unless otherwise specified, all presented approaches are unsupervised. We then briefly discuss available human motion datasets.

A. Human Action Segmentation

Various motion segmentation algorithms at *trajectory level* based on, e.g., zero-velocity crossings (ZVC) [7], principal component analysis [8], or frame-wise-similarities-based clustering [14], were proposed in early literature. These approaches have proven efficient in various applications and have inspired more recent segmentation approaches. For instance, Krüger et al. [15] proposed to consider transition segments between actions based on self-similarity matrices. The proposed algorithms allows the segmentation of cyclic and non-cyclic activities and was evaluated for various input modalities. Krishnan et al. [16] provided a segmentation algorithm that leverages repeated demonstrations to cluster segment endpoints and identify transition states. In contrast to the approaches considered in this paper, this algorithm explicitly requires several demonstrations. Lioutikov et al. [17] proposed to incrementally learn a segmentation and a skill library from demonstrations. An initial over-segmentation was obtained by leveraging different ZVC-like heuristics per task. Tsai et al. [18] proposed a segmentation algorithm for bimanual surgical tasks. Segmentation is achieved by clustering potential segmentation points identified as local extrema and ZVC points of the bimanual distance over time. The resulting spatio-temporal segmentation is further complemented with variance segmentation. Despite its consideration of bimanuality, the aforementioned approach is tailored to specific surgical motions. Klein et al. [19] proposed a segmentation method on trajectory level based on Riemannian geometry, which takes the inertial characteristics of the human body into account. This approach was demonstrated for single-arm gesticulation motions in [19] and is evaluated for bimanual manipulation tasks along with ZVC and acceleration-based segmentation in this paper.

Other approaches segment human motions at *semantic level*. Ma et al. [20] segmented spatio-temporal end-effector trajectories into fine-grained, fuzzy sequences of symbolic vertical and horizontal movements and contacts. However, the approach was mainly evaluated on human locomotion and other whole-body motions. While the approach presented in [20] is tailored to whole-body motion, Wächter et al. [21] proposed to segment human manipulation actions semantically by detecting changes of relations between the human hand and the objects. The authors then extended their approach to a top-down *hierarchical* segmentation method by introducing a trajectory level subsegmentation taking the semantic segments as input [9]. The semantic segments, obtained from contact relations between inflated 3D mesh models, were further subdivided by finding key frames maximizing the difference between Cartesian acceleration profiles. This hierarchic segmentation was evaluated for 13 manipulation action sequences and achieved segmentation closer to manual ground truth than baseline trajectory-level methods. Aksoy et al. [22] first encoded the manipulation task as Semantic Event Chain (SEC), which are patterns of spatial relations between subjects and objects. The motions were segmented based on specified spatial relation changes and subsegmented at local extrema in Cartesian position trajectories. The authors evaluated the action classification based on semantic similarities on 70 partially-bimanual motion capture demonstrations. While the two aforementioned approaches focus on a top-down hierarchy of levels, Gutzeit [23] proposed a supervised bottom-up approach that first segments motions in bell-shaped curves before merging the segments based on the best action classification results on all possible combinations. The approach was evaluated on point-to-point movements and for each hand on dual-arm rotation.

Overall, most of the aforementioned works do not explicitly account for bimanual manipulation actions. Instead, they either consider only the movement of the dominant/active hand in the segmentation, or consider both arms jointly in whole-body segmentation. In this paper, we instead consider each hand individually and evaluate some of the aforementioned existing segmentation algorithms for segmenting bimanual manipulation tasks into individual segments per hand.

B. Human Motion Datasets

Various human motion datasets have been collected in the literature in the close context of multi-modal bimanual recordings of daily household and kitchen activities. Due to the comparatively low cost, there is notably a plethora of single-view video datasets featuring RGB or RGB-D data [6], [24]. However, obtaining accurate data from RGB-D in natural scenarios with multiple, small, and potentially occluded objects is still major area of research. Multi-view RGB-D datasets offer more precise data considering both trajectory-level motions and occurring contacts [25]. However, the most precise and reliable motion data are provided by motion capture systems. Available motion capture datasets such as [26]–[28] either do not provide the precise information about human and object motion, consider only isolated actions, or do not deal with capturing bimanual manipulation actions. We refer the interested reader to our previous work [5] for a comprehensive review of datasets for household tasks published before 2021. More recent motion capture datasets either only consider pick and place tasks instead of complex manipulation [29], hand interaction with a single articulated object [30], very specific actions such as flipping food during grilling [31], or capture only one arm precisely [32].

In contrast, our KIT Bimanual Manipulation Dataset [5] provides precise recordings of human whole-body motions for short bimanual manipulation household actions, e.g., symmetrically rolling dough with both hands or asymmetrically wiping a plate. Each action was executed by two

subjects with a large amount of variations, e.g., type of objects, spatial relations or executed bimanual strategies. This data offers an ideal basis for evaluating segmentation algorithms independent of feature extraction methods such as object and body tracking. In this paper, we address the current limitations of this dataset, i.e., few subjects and short and isolated tasks, by extending it with long manipulation tasks executed by more subjects.

III. Action Segmentation Methods

In this section, we briefly describe the trajectory-level, semantic-level, and hierarchical segmentation algorithms that are applied on the Extended KIT Bimanual Manipulation Dataset in Section V. Note that all algorithms are based on previous works. The exact formulation, the parameter variables, or the application, if different, is specified. All methods perform segmentation of each human demonstration independently, i.e., no inter-motion features are considered.

A. Trajectory Level Segmentation

In this paper, we consider three algorithms for segmentation at trajectory level. These algorithms are based on zero-velocity crossing [7], acceleration profiles [9], and geodesic segmentation [19], as explained next.

- *Zero-Velocity Crossing (ZVC).* Motions are segmented when at least $n \leq N$ dimensions within a sliding window of size t cross zero velocity [7]. Segments composed of minor movements or oscillations are discarded by ignoring zero-velocity crossings in a given dimension if the average distance to the mean within the sliding window is smaller than a threshold Δ_μ.
- *Acceleration Profile (ACC).* Segmentation points are extracted based on maximizing the difference of the trajectory before and after a segmentation frame. This is done in an iterative way where for each segment the frame with the highest quality measure $\hat{q}_{best}$ is selected as an additional segmentation point if $\hat{q}_{best} > \lambda$ and the resulting segments have a size larger than l_{min}. The heuristic for the quality measure is based on the accelerations in all dimensions considering both their peak-to-peak amplitude and the curve length approximated by frame-wise differences. The quality measure at frame i is computed based on a window $[i - \frac{w}{2}, i + \frac{w}{2}]$. The precise definition of the heuristic can be found in [9].
- *Geodesic (GEO).* The human configuration space is viewed as a Riemannian manifold endowed with the kinetic-energy metric [19]. Human motions are piecewise geodesic with respect to this manifold and motions are segmented at transitions between geodesics. A transition is detected if the angle θ between the current velocity $\dot{\boldsymbol{q}}_t$ and the parallel-transported initial velocity of the segment is larger than a threshold Δ_θ. Note that we disregard segmentation points for which $||\dot{\boldsymbol{q}}_t|| < \Delta_{\dot{q}}$ as they correspond to minor oscillations in the motion.

Notice that ZVC and GEO tend to over-segment the motion. Therefore, we further filter the final segmentation points as follows. Successive segmentation points are combined into an interval $I = [a, b]$ if the point-to-point time difference is smaller than a threshold Δ_s. We consider the interval boundaries a and b as segmentation points if $b - a > \Delta_w$, and otherwise a mean segmentation point $c = \frac{a+b}{2}$.

B. Semantic Level Segmentation

Semantic segmentation algorithms such as [9], [22] determine segments based on changes of spatial relations between the human and objects in the environment. Such segments are thus intrinsically endowed with a semantic interpretation. In this paper, we consider the semantic segmentation approach of [9] and extend it to bimanual manipulation actions. To handle bimanuality, each contact occurring in the scene, i.e., each segmentation point, needs to be assigned to one of the hands. To solve this problem, we propose to first determine the grasped object based on the contact relation, so that grasping or releasing an object adds a segmentation point for the corresponding hand. The object in each hand is determined by contact using a distance threshold Δ_{ho}. Notice that we only consider contacts between the hand and manipulation objects, but ignore contacts with the environment or with the other hand. Moreover, in order to filter out small inaccuracies, changes in contact relations are ignored if they occur only for less than a threshold Δ_c. An additional distance threshold Δ_{oo} is used to detect changes in object-object contact relations. Moreover, in order to only consider the most relevant contacts, we count the number of objects that are in contact with the grasped object and only add a segmentation point if the number of contact relation changes to exactly zero or one.

C. Hierarchic Segmentation

While semantic segmentation ensures to produce somewhat meaningful segments, it lacks the granularity to extract all essential parts of a task. Hierarchical segmentation methods tackle this issue by applying a trajectory level subsegmentation on segments extracted by semantic methods. In this paper, we evaluate the hierarchical segmentation algorithm of [9], whose upper level leverages the semantic segmentation summarized in Section III-B. For an extensive evaluation, we consider the different trajectory-level segmentations presented in Section III-A for the lower level.

IV. Extended KIT Bimanual Manipulation Dataset

This paper aims at evaluating the action segmentation methods described in Section III on bimanual manipulation tasks in household scenarios. To do so, we leverage our recently-published KIT Bimanual Manipulation Dataset [5], which offers multi-modal recordings of isolated household actions with a high variance in their execution. The dataset includes two subjects and rather short, scripted recordings. However, handling inter-subject variations and transitions between different actions are major challenges in action segmentation. Thus, such variations and transitions should be considered when comparing segmentation algorithms.

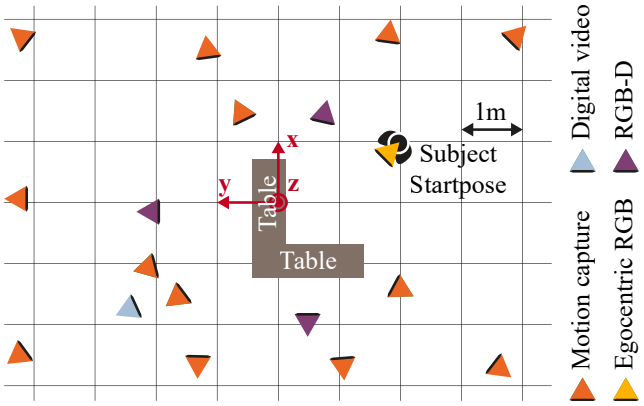


Fig. 1: Setup and positions of the cameras.

In this section, we extend the KIT Bimanual Manipulation Dataset with recordings of the same two and four more subjects performing sequences of the same bimanual household actions in randomized scenes. An overview of the type and number actions in both datasets is given by Table I. For comparability, we maintained the same multi-modal sensor setup and objects as in [5]. The extended dataset results in about 176 minutes of recorded motions. This study was approved by the ethics committee of the Karlsruhe Institute of Technology, Karlsruhe, Germany. The participants gave their written informed consent before the experiments that the data may be made publicly available for research purposes in the KIT Whole-Body Human Motion Database [33].

A. Setup and Recording Procedure

As in our previous work [5], the bimanual human motions are captured by means of an optical motion capture system, two data gloves, three inertial measurement units, three RGB-D cameras, and an egocentric RGB camera. Our sensor setup is visualized in Figure 1. While the same objects as for the previous dataset are used, an extra table is added to the scene, providing more space for manipulation and at the same time forming a kitchen corner. The positions of the digital video, RGB-D, and motion capture cameras were slightly adjusted to accommodate this change. Each recording begins with the subject standing in a T-pose about one to two meters from the tables (see Figure 2).

Six healthy, right-handed[1] subjects (three male, three female) performed sequences of manipulation actions within three daily household scenarios, namely preparing a meal, preparing a pie and cleaning up. While the subjects were informed about the overall goal to achieve, the execution of the actions and their order was left to their own discretion. More information about the scenarios can be found in the description of the recordings in our motion database. Altogether, each subject repeated every scenario five times, resulting in 90 recordings for a total of about 104 minutes. After each repetition, the positions of objects on and next to the table were randomly changed by the instructor. Figure 2 shows an exemplary scene for preparing a meal.

[1]Data from left-handed subjects will be collected as future work.

Fig. 2: Subject in T-pose in front of a random scene for the task preparing a meal.

B. Mapping and Representation

The recorded multi-modal sensor data is converted to the Master Motor Map (MMM) data format [33] which is a unified representation for whole-body human motion data. Thereby, the marker-based motion capture is mapped to the MMM reference model -— a reference model of the human body including statistical, kinematic, and dynamic properties -— and to modelled or 3D-scanned mesh models [5] of all objects. The mapping is achieved by minimizing the error between recorded and virtual marker positions obtained on the model with forward kinematics [33]. This is solved using a sequential quadratic programming optimization via NLopt [34] including additional criteria inspired by [35] to reduce the joint velocity, acceleration, and jerk [19]. For each frame, we can then retrieve the joint configuration of the human body, the 6D pose of the human root and of objects, as well as the 6D pose of any segment (e.g., the hand) via forward kinematics.

C. Bimanual Segmentation and Annotation

The data is manually segmented and annotated for each hand as for the previous dataset. As shown in Table I, these annotations include manipulation phases (approach, lift, place, retreat), supporting actions (hold, move), and manipulation actions (e.g., cut, stir), which are further subsegmented. Two new action labels were added in the extended dataset, namely regrasping an object and shaking, e.g., the whisk after stirring. Moreover, actions were further annotated with failure cases, i.e., the drop and slip of an object, as well as planning failures, which happened when an action was partially performed before the intent was changed, or when an incorrect action was accidentally performed, e.g., turning the lid in the wrong direction when opening/closing a bottle.

V. Evaluation

In this section, we evaluate the segmentation algorithms described in Section III on the KIT Bimanual Manipulation Dataset [5] and its extension presented in Section IV.

A. Data Processing and Parameter Selection

We apply the segmentation algorithms on the precise human kinematic and 6D object pose data obtained via motion capture. Although the algorithms could also be executed on

TABLE I: Manually-segmented actions of the Extended KIT Bimanual Manipulation Dataset. The top level refers to manipulation phases, supporting, and manipulation actions and the bottom level corresponds to subsegmentation thereof.

	Level	Top				Bottom			
	Dataset	Original [5]		Extension		Original [5]		Extension	
	Hand	Left	Right	Left	Right	Left	Right	Left	Right
Grasp / Support	**Approach**	615	599	557	597	0	0	0	0
	Hold	486	123	477	270	0	0	4	1
	Lift	446	464	350	318	0	8	0	0
	Move	181	73	320	238	175	325	442	880
	Place	502	537	447	476	0	0	0	0
	Regrasp	0	0	92	47	0	0	0	0
	Retreat	665	628	530	558	0	0	0	0
Manipulation	**Close**	36	0	37	14	84	0	102	66
	Cut	0	42	0	31	0	126	0	117
	Mix	18	18	33	33	54	54	145	146
	Open	24	0	33	32	51	0	120	102
	Peel	0	12	0	57	0	36	0	312
	Pour	18	83	30	54	18	177	33	114
	RollOut	25	24	65	65	109	109	361	361
	Scoop	18	113	31	90	18	113	31	90
	Shake	0	0	2	42	0	0	0	0
	Stir	0	78	0	32	0	235	0	317
	Sweep	30	30	33	33	90	90	162	162
	Transfer	0	72	6	85	18	90	32	112
	Wipe	0	54	3	74	0	190	15	544

noisy data or data from other data sources, considering precise motion data avoids evaluating the approaches based on the underlying feature extraction method, e.g., in the context of vision models. The velocities and accelerations of the joints and hands are computed as the analytical derivatives of a second-order approximation of the joint angles over time obtained with a Savitzky–Golay filter (window-length of 21 samples and order 3). Contacts are computed between inflated 3D mesh models of the human hand and objects. Specifically, the human palm and proximal and intermediate phalanges of the index finger and thumb are considered for contacts. For hollow objects, e.g. cups and bowls, the convex hull is used as collision model.

Many segmentation heuristics can be applied on trajectories in either task space or joint space. For a thorough comparison, we apply ZVC and ACC on *(i)* the Cartesian positions of the tool center of each hand (hereinafter denoted as $\text{ZVC}_{\boldsymbol{x}}$, $\text{ACC}_{\boldsymbol{x}}$), and *(ii)* the 7 joint angles of each arm (denoted as $\text{ZVC}_{\boldsymbol{q}}$, $\text{ACC}_{\boldsymbol{q}}$). The parameters of the different segmentation algorithms are fixed via a grid search on a subset of 12 recordings of different actions of the same subject from [5]. For trajectory-level segmentation, we retain the parameters maximizing the F_1 score. For semantic-level, we select parameters that lead to high precision, since a higher level segmentation should find meaningful segmentation points. Table III shows the different segmentation methods used for our evaluation along with their corresponding optimized parameters. The computation of the aforementioned evaluation metrics (F_1 score and precision) is detailed next.

B. Evaluation Metric

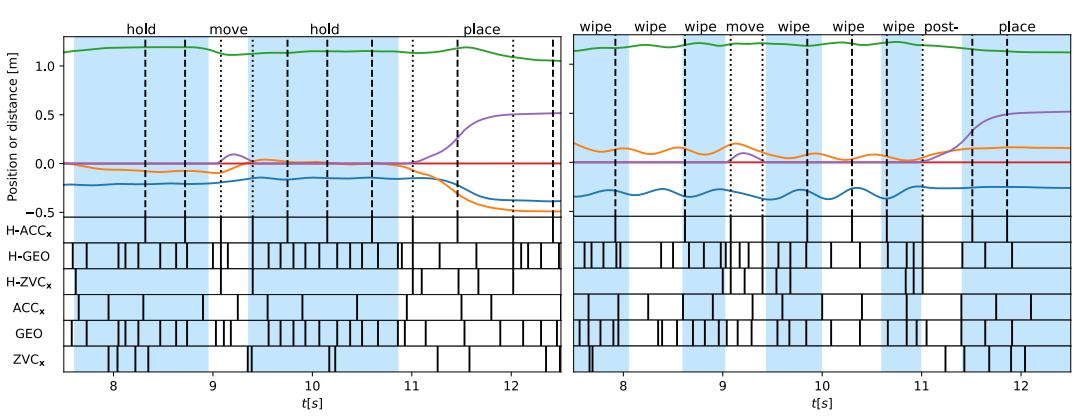
We evaluate the quality of the considered segmentation algorithms both qualitatively and quantitatively in terms of *precision*, *recall*, and F_1 scores. To do so, the manual annotations provided along with the dataset (see Section IV-C) are considered as ground truth. We leverage the Integrated Kernel (InK) approach [36] to meaningfully compute the aforementioned quantitative quality measures, while taking temporal distances into account. Namely, InK computes the amount of true/false positives/negatives based on the integrals of two segmentation functions of opposed sign, which represent the segmentation points of the ground truth and evaluated segmentation with kernel functions. Here, we choose a Gaussian kernel function with a variance $\sigma = 111.11\,\text{ms}$, so that about $99.7\,\%$ of the area of the Gaussian distribution lies in the range of $\pm 3\sigma = 333\,\text{ms}$. For our evaluation, we merge the two granularity levels of segmentation of the manual annotations to a single level of segmentation.

TABLE II: Segmentation algorithms with their parameters.

Name	**Parameters**			
$\text{ACC}_{\boldsymbol{q}}$	$l_{min} = 250\,\text{ms}$	$\lambda = 5$	$w = 400\,\text{ms}$	$z = 0.5$
$\text{ACC}_{\boldsymbol{x}}$	$l_{min} = 300\,\text{ms}$	$\lambda = 1$	$w = 250\,\text{ms}$	$z = 0.5$
GEO	$\Delta_\theta = 0.8$	$\Delta_{\dot{q}} = 0.3$		
$\text{ZVC}_{\boldsymbol{q}}$	$n = 2$	$\Delta_\mu = 0.005$	$t = 200\,\text{ms}$	
$\text{ZVC}_{\boldsymbol{x}}$	$n = 1$	$\Delta_\mu = 0.001$	$t = 200\,\text{ms}$	
SEM	$\Delta_{ho} = 0.01\,\text{m}$	$\Delta_{oo} = 0.005\,\text{m}$	$\Delta_c = 50\,\text{ms}$	
H-$\text{ACC}_{\boldsymbol{q}}$	$l_{min} = 250\,\text{ms}$	$\lambda = 10$	$w = 350\,\text{ms}$	$z = 0.5$
H-$\text{ACC}_{\boldsymbol{x}}$	$l_{min} = 350\,\text{ms}$	$\lambda = 5$	$w = 200\,\text{ms}$	$z = 0.5$
H-GEO	$\Delta_\theta = 0.8$	$\Delta_{\dot{q}} = 0.3$		
H-$\text{ZVC}_{\boldsymbol{q}}$	$n = 1$	$\Delta_\mu = 0.02$	$t = 250\,\text{ms}$	
H-$\text{ZVC}_{\boldsymbol{x}}$	$n = 1$	$\Delta_\mu = 0.01$	$t = 250\,\text{ms}$	
All	$\Delta_s = 50\,\text{ms}$	$\Delta_w = 300\,\text{ms}$		

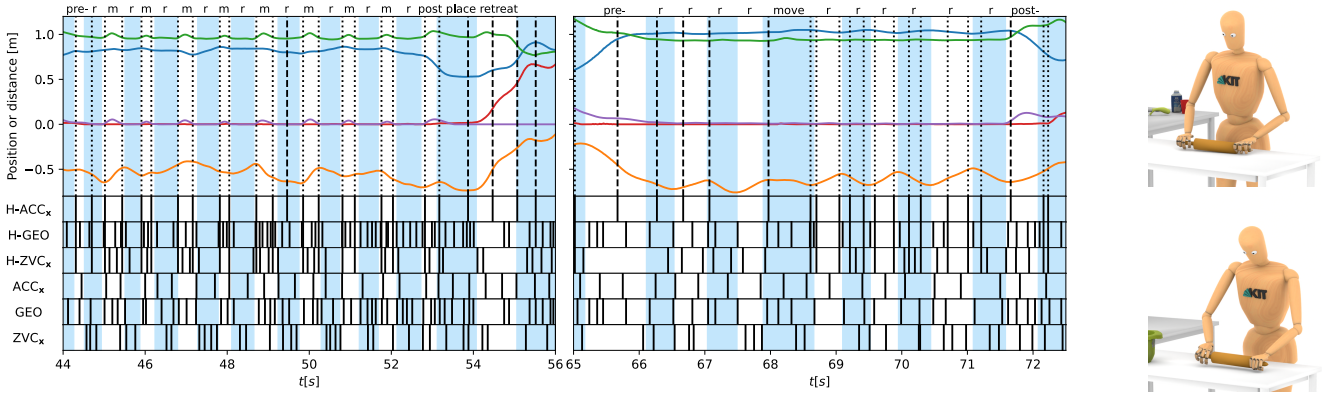
C. Results

Figure 3 shows the segmentation points obtained by different segmentation algorithms for each hand along the wipe action within the cleaning up task. For clarity of presentation, we only display three trajectory-level ($\text{ACC}_{\boldsymbol{x}}$, GEO, $\text{ZVC}_{\boldsymbol{x}}$) and three hierarchical (H-$\text{ACC}_{\boldsymbol{x}}$,H-GEO,H-$\text{ZVC}_{\boldsymbol{x}}$) segmentation algorithms. We observe that $\text{ZVC}_{\boldsymbol{x}}$ and to some extent $\text{ACC}_{\boldsymbol{x}}$ fail to identify many segmentation points given by the manual annotations and result in few false positive, i.e., segmentation points that were not manually annotated. Instead, GEO matches almost all manual segmentation points at the expense of generating more false positive. This can be explained by the fact that the geodesic segmentation explicitly consider the dynamics underlying human motions. In other words, each manually-annotated segment corresponds to a dynamic motion that is naturally identified with a sequence of simple dynamic motion units detected via GEO. Therefore, GEO consistently segments motions at a lower level than the manual annotations. Table III presents a quantitative evaluation of the different segmentation algorithms. In accordance with Figure 3, ZVC and ACC result in similar precision and recall with slightly higher scores for ACC. Moreover, GEO reaches a high recall at the expense of a low precision due to its natural tendency to over-segment the motion in comparison to the manual annotations.



(a) Left hand segmentations (b) Right hand segmentations (c) $t = \{9.5, 10\}$s

Fig. 3: Bimanual segmentation for the wipe action within the cleaning up task for the subject 1480. *(a)-(b)* Comparisons of the segmentation points ( ) obtained with different algorithms. The manual annotations of the dataset ( ) are considered as ground truth. The *top* panels differentiate between the semantic segmentation ( ) and the trajectory subsegmentation ACC$_{\boldsymbol{x}}$ ( ) of the hierarchical segmentation H-ACC$_{\boldsymbol{x}}$. The hand trajectories (x_1 —, x_2 —, x_3 —) and the distances between the hand and *(a)* the cutting board or *(b)* the sponge (—) and between the sponge and the cutting board (—) are also depicted. *(c)* Snapshots of the task mapped onto the MMM model.



(a) Left hand segmentation for subject 1723 (b) Left hand segmentation for subject 1747 (c) $\{46.5, 67.5\}$s

Fig. 4: Bimanual segmentation for the rollout action within the preparing a pie task for the two different subjects. r and m denotes rollout and move actions throughout the task, respectively. *(a)-(b)* Comparisons of the segmentation points ( ) obtained with different algorithms. The manual annotations of the dataset ( ) are considered as ground truth. The *top* panels differentiate between the semantic segmentation ( ) and the trajectory subsegmentation ( ) of the hierarchical segmentation H-ACC$_{\boldsymbol{x}}$. The hand trajectories (x_1 —, x_2 —, x_3 —) and the distances between the hand and the rolling pin (—) and between the rolling pin and the table (—) are also depicted. *(c)* Snapshots of the task mapped onto the MMM model. The *top* and *bottom* models display the motion of subject 1723 and 1747, respectively.

As shown in the top panel of Figure 3, the semantic segmentation (SEM) successfully distinguishes the start and end of the successive wiping motions via the contact changes. This is also observed in Table III as SEM results in the highest precision score. Notice that this is the desired behavior for the high level of a hierarchical segmentation. In general, the subsegmentations of the hierarchical approaches H-ACC$_{\boldsymbol{x}}$, H-GEO, H-ZVC$_{\boldsymbol{x}}$ behave similarly as the corresponding algorithms when used at trajectory levels. Namely, H-ACC$_{\boldsymbol{x}}$, H-ZVC$_{\boldsymbol{x}}$ still fail to identify some segmentation points, while H-GEO segments motions at a lower level than the manual annotation. However, the identified segmentation points are not always identical, especially for H-ACC$_{\boldsymbol{x}}$ vs ACC$_{\boldsymbol{x}}$ and H-ZVC$_{\boldsymbol{x}}$ vs ZVC$_{\boldsymbol{x}}$. Moreover, all algorithms segment the successive wiping actions similarly for the dominant (right) and non-dominant (left) hand, although the wiping was primarily achieved with the dominant hand. This is explained by the fact that the non-dominant hand still performed subtle circular motions while holding the cutting board. Notice that this issue arises for various asymmetric motions, for which the supporting actions of the non-dominant hand were annotated as a single hold action. This is also observed in

TABLE III: Evaluation of segmentation algorithms on the Extended KIT Bimanual Manipulation Dataset. The manual annotations provided in the dataset are considered as ground truth.

Dataset	Score	Arm	ACC$_q$	ACC$_x$	GEO	ZVC$_q$	ZVC$_x$	SEM	H-ACC$_q$	H-ACC$_x$	H-GEO	H-ZVC$_q$	H-ZVC$_x$
Original [5]	$F1$	Right	**0.54**	**0.55**	0.40	0.46	0.47	0.42	**0.54**	**0.53**	0.37	0.44	0.46
		Left	**0.48**	**0.48**	0.34	0.41	0.44	0.37	**0.46**	**0.46**	0.32	0.39	0.39
	$Precision$	Right	0.41	0.44	0.26	0.35	0.36	**0.52**	0.41	0.45	0.23	0.31	0.37
		Left	0.34	0.36	0.21	0.29	0.32	**0.42**	0.33	0.37	0.19	0.26	0.30
	$Recall$	Right	0.80	0.72	**0.90**	0.65	0.66	0.36	0.81	0.66	**0.93**	0.73	0.60
		Left	0.79	0.71	**0.88**	0.69	0.73	0.34	0.79	0.62	**0.91**	0.74	0.55
Extension	$F1$	Right	**0.54**	**0.55**	0.46	0.39	0.39	0.44	**0.54**	**0.54**	0.40	0.45	0.49
		Left	**0.40**	**0.41**	0.31	0.30	0.31	0.39	**0.41**	**0.41**	0.28	0.34	0.42
	$Precision$	Right	0.46	0.48	0.30	0.33	0.33	**0.52**	0.44	0.48	0.25	0.33	0.40
		Left	0.28	0.30	0.18	0.21	0.21	**0.39**	0.28	0.32	0.17	0.23	0.31
	$Recall$	Right	0.67	0.64	**0.92**	0.50	0.48	0.38	0.73	0.63	**0.94**	0.70	0.63
		Left	0.66	0.63	**0.89**	0.55	0.60	0.39	0.71	0.57	**0.93**	0.68	0.68

Table III, where the precision and F_1 score of the dominant, i.e., right[2], hand are higher than for the non-dominant one.

Figure 4 shows the segmentation points obtained for two different subjects during the rollout action within the preparing a pie task. As opposed to the asymmetric wipe action, both hands synchronously perform the same motions during rollout and we only display the left hand segmentation. We observe that all approaches generally display similar characteristics as in Figure 3, thus demonstrating consistency throughout different actions. Moreover, Table III shows that the different segmentation algorithms achieve consistent performances on both the original dataset [5] and its extension (Section IV). Importantly, we also observe consistent behavior of the different segmentation approaches across subjects (see Figure 4a vs Figure 4b). During the rollout action, SEM sometimes overlooks the manually-annotated segmentation points. This is due to the fact that contacts between the rolling pin and the table are not detected when the dough is too thick. However, these segments are recovered by the segmentation at trajectory level and thus by the subsegmentation of the hierarchical algorithms.

Overall, we observe that the hierarchical segmentation benefits from the semantic segmentation, which often matches the manual annotations. Therefore, we generally observe a slight increase in the recall of hierarchical algorithms compared to their trajectory-level counterpart. Interestingly, all segmentation approaches achieve similar F_1, precision and recall scores on both datasets. Finally, we observe only minor differences on the performances of the ACC and ZVC when applied at the hand position or at the joint trajectory level.

VI. Discussion

This paper presented a detailed evaluation of various segmentation algorithms at trajectory and semantic levels, as well as hierarchical segmentation algorithms on bimanual manipulation tasks. To handle bimanuality, we proposed to segment the motion of each hand individually, and extended a semantic segmentation approach used within hierarchical algorithms to handle bimanual actions. All algorithms were evaluated on the KIT Bimanual Manipulation Dataset [5], which was further extended in this paper to include long bimanual manipulation tasks recorded with various subjects.

Our evaluation showed that the considered segmentation algorithms display similar characteristics across short and long sequences of actions, and subjects. Moreover, it confirmed that hierarchical segmentation algorithms benefit from the segmentation at semantic and trajectory level to generate meaningful segmentation accounting for additional variations within hand or joint trajectories. In addition, we believe that geodesic (sub)segmentation presents several advantages compared to the widely-used (sub)segmentation based on ZVC and acceleration profiles. As it intrinsically account for the dynamic properties of the human body, the resulting segments correspond to intrinsic human motion units, which may later be leveraged to built human-like libraries of dynamic primitives.

Importantly, our evaluation suggests that taking the role of the hands into account would be beneficial when segmenting bimanual manipulation tasks, especially when hands have asymmetric and possibly varying roles throughout the task. In particular, we believe that rule-based approaches such as [2] may be combined with segmentation algorithms to infer the role of the hands and adapt the parameters of trajectory-level (sub)segmentation accordingly. We plan to investigate such algorithms in our future work.

It is important to emphasize that inaccuracies may occur in the manual annotations, which lead to a decrease of the quantitative performance achieved by the different segmentation methods. For instance, the semantic segmentation was often slightly shifted compared to the manual annotations in Figure 4. Moreover, our evaluation was conducted on motion capture data with relatively low noise, except for the motion and mesh of the manipulated real food items, e.g. cucumber or eggplant, which could only be reconstructed semi-accurately. Noisy setups may require adaptation on segmentation approaches, as well as additional evaluations.

Overall, bimanual action segmentation is the first step

[2]All subjects recorded within the extended dataset are right-handed.

towards building libraries of bimanual movement primitives and learning task models for bimanual manipulation. Constructing such libraries and task models still remains a large research topic. As future work, we plan to tackle some of these challenges by representing extracted motion segments with suitable uni- and bimanual motion representations.

ACKNOWLEDGMENT

The authors would like to acknowledge the help and support of the motion capture students (Saghar Samaei, Mitat Hoxha, Kevin Liang, Adnan Ügür, and Sven Henkel) of the H^2T lab who contributed to the presented dataset by assisting during the recording sessions and the post-processing of the motion capture recordings.

REFERENCES

[1] A. Billard, S. Calinon, R. Dillmann, and S. Schaal, *Robot Programming by Demonstration*. Springer Berlin Heidelberg, 2008, pp. 1371–1394.

[2] F. Krebs and T. Asfour, "A bimanual manipulation taxonomy," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 11 031–11 038, 2022.

[3] A. Billard and D. Kragic, "Trends and challenges in robot manipulation," *Science*, vol. 364, no. 6446, 2019.

[4] F. Meier, E. Theodorou, F. Stulp, and S. Schaal, "Movement segmentation using a primitive library," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2011, pp. 3407–3412.

[5] F. Krebs, A. Meixner, I. Patzer, and T. Asfour, "The KIT bimanual manipulation dataset," in *IEEE/RAS Intl. Conf. on Humanoid Robots (Humanoids)*, 2021, pp. 499–506.

[6] C. R. G. Dreher, M. Wächter, and T. Asfour, "Learning object-action relations from bimanual human demonstration using graph networks," *IEEE Robotics and Automation Letters*, vol. 5, no. 1, pp. 187–194, 2020.

[7] A. Fod, M. Mataric, and O. Jenkins, "Automated derivation of primitives for movement classification," *Autonomous Robots*, vol. 12, pp. 39–54, 2003.

[8] J. Barbič, A. Safonova, J.-Y. Pan, C. Faloutsos, J. K. Hodgins, and N. S. Pollard, "Segmenting motion capture data into distinct behaviors," in *Proceedings of Graphics Interface*, 2004, p. 185–194.

[9] M. Wächter and T. Asfour, "Hierarchical segmentation of manipulation actions based on object relations and motion characteristics," in *IEEE Intl. Conf. on Advanced Robotics (ICAR)*, 2015, pp. 549–556.

[10] H. Xing and D. Burschka, "Understanding spatio-temporal relations in human-object interaction using pyramid graph convolutional network," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2022, pp. 5195–5201.

[11] E. G. Herrero, J. Ho, and O. Khatib, "Understanding and segmenting human demonstrations into reusable compliant primitives," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2021, pp. 9437–9444.

[12] T. Nakamura, T. Nagai, D. Mochihashi, I. Kobayashi, H. Asoh, and M. Kaneko, "Segmenting continuous motions with hidden semi-Markov models and Gaussian processes," *Frontiers in Neurorobotics*, vol. 11, 2017.

[13] M. Nagano, T. Nakamura, T. Nagai, D. Mochihashi, I. Kobayashi, and W. Takano, "High-dimensional Motion Segmentation by Variational Autoencoder and Gaussian Processes," Nov. 2019, pp. 105–111, iSSN: 2153-0866.

[14] F. Zhou, F. De la Torre, and J. K. Hodgins, "Hierarchical aligned cluster analysis for temporal clustering of human motion," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 582–596, 2013.

[15] B. Krüger, A. Vögele, T. Willig, A. Yao, R. Klein, and A. Weber, "Efficient unsupervised temporal segmentation of motion data," *IEEE Trans. on Multimedia*, vol. 19, no. 4, pp. 797–812, 2017.

[16] S. Krishnan, A. Garg, S. Patil, C. Lea, G. Hager, P. Abbeel, and K. Goldberg, "Transition state clustering: Unsupervised surgical trajectory segmentation for robot learning," *Intl. Journal of Robotics Research*, vol. 36, no. 13-14, pp. 1595–1618, 2017.

[17] R. Lioutikov, G. Neumann, G. Maeda, and J. Peters, "Learning movement primitive libraries through probabilistic segmentation," *Intl. Journal of Robotics Research*, vol. 36, no. 8, pp. 879–894, 2017.

[18] Y.-Y. Tsai, Y. Guo, and G.-Z. Yang, "Unsupervised task segmentation approach for bimanual surgical tasks using spatiotemporal and variance properties," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2019, pp. 1–7.

[19] H. Klein, N. Jaquier, A. Meixner, and T. Asfour, "A riemannian take on human motion analysis and retargeting," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2022, pp. 5210–5217.

[20] H. Ma, Z. Yang, and H. Liu, "Fine-grained unsupervised temporal action segmentation and distributed representation for skeleton-based human motion analysis," *IEEE Trans. on Cybernetics*, vol. 52, no. 12, pp. 13 411–13 424, 2022.

[21] 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 Intl. Conf. on Humanoid Robots (Humanoids)*, 2013, pp. 189–195.

[22] E. E. Aksoy, Y. Zhou, M. Wächter, and T. Asfour, "Enriched manipulation action semantics for robot execution of time constrained tasks," in *IEEE/RAS Intl. Conf. on Humanoid Robots (Humanoids)*, 2016, pp. 109–116.

[23] L. Gutzeit, "Hierarchical segmentation of human manipulation movements," in *Intl. Conf. on Pattern Recognition (ICPR)*, 2022, pp. 2742–2748.

[24] J. Jang, D. Kim, C. Park, M. Jang, J. Lee, and J. Kim, "ETRI-Activity3D: A large-scale rgb-d dataset for robots to recognize daily activities of the elderly," *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pp. 10 990–10 997, 2020.

[25] B. L. Bhatnagar, X. Xie, I. A. Petrov, C. Sminchisescu, C. Theobalt, and G. Pons-Moll, "BEHAVE: Dataset and method for tracking human object interactions," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, 2022, pp. 15 914–15 925.

[26] F. de la Torre Frade, J. K. Hodgins, A. W. Bargteil, X. M. Artal, J. C. Macey, A. C. I. Castells, and J. Beltran, "Guide to the Carnegie Mellon University multimodal activity (CMU-MMAC) database," Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMU-RI-TR-08-22, April 2008.

[27] Y. Huang and Y. Sun, "A dataset of daily interactive manipulation," *Intl. Journal of Robotics Research*, vol. 38, no. 8, pp. 879–886, 2019.

[28] O. Taheri, N. Ghorbani, M. J. Black, and D. Tzionas, "GRAB: A dataset of whole-body human grasping of objects," in *European Conf. on Computer Vision*, 2020, pp. 581–600.

[29] P. Kratzer, S. Bihlmaier, N. B. Midlagajni, R. Prakash, M. Toussaint, and J. Mainprice, "MoGaze: A dataset of full-body motions that includes workspace geometry and eye-gaze," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 367–373, 2021.

[30] Z. Fan, O. Taheri, D. Tzionas, M. Kocabas, M. Kaufmann, M. J. Black, and O. Hilliges, "ARCTIC: A dataset for dexterous bimanual hand-object manipulation," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[31] D. Pereira, Y. De Pra, E. Tiberi, V. Monaco, P. Dario, and G. Ciuti, "Flipping food during grilling tasks, a dataset of utensils kinematics and dynamics, food pose and subject gaze," *Scientific Data*, vol. 9, 2022.

[32] E. Nicora, G. Goyal, N. Noceti, A. Vignolo, A. Sciutti, and F. Odone, "The MoCA dataset, kinematic and multi-view visual streams of fine-grained cooking actions," *Scientific Data*, vol. 7, 2020.

[33] C. Mandery, Ö. Terlemez, M. Do, N. Vahrenkamp, and T. Asfour, "Unifying representations and large-scale whole-body motion databases for studying human motion," *IEEE Trans. on Robotics*, vol. 32, no. 4, pp. 796–809, 2016.

[34] S. G. Johnson, *The NLopt nonlinear-optimization package*, 2011. [Online]. Available: http://ab-initio.mit.edu/nlopt

[35] D. Rakita, B. Mutlu, and M. Gleicher, "RelaxedIK: Real-time synthesis of accurate and feasible robot arm motion," in *Robotics: Science and Systems (R:SS)*, 2018.

[36] C. R. G. Dreher, N. Kulp, C. Mandery, M. Wächter, and T. Asfour, "A framework for evaluating motion segmentation algorithms," in *IEEE/RAS Intl. Conf. on Humanoid Robots (Humanoids)*, 2017, pp. 83–90.