The KIT Bimanual Manipulation Dataset

Franziska Krebs*, Andre Meixner*, Isabel Patzer and Tamim Asfour

***Abstract*— Learning models of bimanual manipulation tasks from human demonstration requires capturing human body and hand motions, as well as the objects involved in the demonstration, to provide all the information needed for learning manipulation task models on symbolic and subsymbolic level. We provide a new multi-modal dataset of bimanual manipulation actions consisting of accurate human whole-body motion data, full configuration of both hands, and the 6D pose and trajectories of all objects involved in the task. The data is collected using five different sensor systems: a motion capture system, two data gloves, three RGB-D cameras, a head-mounted egocentric camera and three inertial measurement units (IMUs). The dataset includes 12 actions of bimanual daily household activities performed by two healthy subjects with a large number of intra-action variations and three repetitions of each action variation, resulting in 588 recorded demonstrations. A total of 21 household items are used to perform the various actions. In addition to the data collection, we developed tools and methods for the standardized representation and organization of multi-modal sensor data in large-scale human motion databases. We extended our Master Motor Map (MMM) framework to allow the mapping of collected demonstrations to a reference model of the human body as well as the segmentation and annotation of recorded manipulation tasks. The dataset includes raw sensor data, normalized data in the MMM format and annotations, and is made publicly available in the KIT Whole-Body Human Motion Database.**

I. Introduction

Robot programming by demonstration (PbD) is a promising and effective approach for teaching robots new skills in an intuitive way and by non-expert users [1]. While PbD in general has been an area of extensive research for decades, learning bimanual manipulation tasks from human demonstration is a largely underdeveloped area and remains a challenging task [2]. This is due to the fact, that bimanual manipulation actions are not simply the sum of two unimanual actions because temporal and spatial coordination as well as potential interactions between the hands must be taken into account. In the context of learning motion primitives from demonstration ([3], [4]), kinesthetic teaching is often applied to generate data for learning on the target robot system while avoiding the correspondence problem and motion retargeting between different embodiments [5]. However, kinesthetic teaching does not provide the information about temporal and spatial relations between hands and objects to be manipulated, which is needed for bimanual coordination and goal-directed adaptation of learned bimanual actions.

Fig. 1: Left: Example of a bimanual action (*transfer*). Right: Subject in full body suit with markers and sensors. The multi-modal sensor setup used to collect the data is partially visible.

We consider such object-hand and hand-hand relations key constraints for learning semantic task models for bimanual manipulation and their goal-directed adaptation and execution. Such semantic task models should encode i) symbolic task information such as actions with their preconditions and effects, spatial and temporal relations between hands and objects as well as ii) subsymbolic, sensorimotor information needed for learning bimanual movement primitives from e. g. position trajectories, force profiles and embodiment specific parameters.

Providing datasets containing all the information needed for learning such task models of bimanual manipulation from human demonstration, their mapping to and execution on different robots is a challenging task that requires significant efforts ranging from a systematic recording of human and object motion data, unified representation of collected data to providing methods and tools for processing and interpretation of the data. Such datasets would lead to considerable progress in the area of learning bimanual tasks by facilitating several research directions and contributing to the reproducibility of research results in this area. To the best of our knowledge, there is no comprehensive dataset that explicitly accounts for the bimanuality of actions while providing all the information needed to extract semantic and sensorimotor information from human demonstration in terms of accurate multi-modal recordings of human whole-body motion, full human hand configuration as well as pose and motions of all objects involved in the task.

Other approaches address the question of collecting data of human demonstrations for learning by capturing interactions of humans with virtual environments ([6]–[8]) and their

* The first two authors contributed equally to this work.

This work has been supported by the German Federal Ministry of Education and Research (BMBF) under the project OML (01IS18040A).

The authors are with the Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Germany {franziska.krebs, andre.meixner,isabel.patzer,asfour}@kit.edu

physical simulation or by using computer vision methods to synthetically generate photo-realistic renderings of human poses based on image and human motion capture data [9]. While both approaches are promising, they are limited in terms of representing realistic physical interactions with objects. This is due to the fact that data generated in virtual reality strongly relies on the accuracy of the underlying physical simulation and data generated based on image data and motion capture data such as in [9] rarely includes object information. Both approaches can highly benefit from accurate motion capture data as ground-truth for evaluation or even rely on such motion capture data as in e. g. [9].

Our Contribution: We present a new multi-modal dataset of bimanual manipulation actions consisting of i) accurate human whole-body motion data, ii) full configuration of both hands and iii) 6D pose and motion of all objects involved in the task. The data is collected using five different sensor systems: i) VICON motion capture system to record whole-body human and object motion with high accuracy at the trajectory level, ii) two data gloves that provide finger joint trajectories of both hands, iii) three RGB-D cameras that provide different perspectives on the scene, iv) a head-mounted egocentric camera to capture the subject's field of view and v) three inertial measurement units (IMUs) attached to the human body to provide additional information and investigate the potential of minimalistic wearable sensor setups in the context of action recognition. The dataset includes 12 bimanual actions of daily household activities performed by two subjects with a large number of variations within actions. In total, 21 household objects are used for the execution of the different actions. An excerpt of our motion recordings can be seen in Figure 1. The data is further segmented and annotated to facilitate future use and research in the area of learning semantic task models on symbolic and subsymbolic level of bimanual manipulation. Following a unifying approach for the representation and organization of large-scale human motion databases, we extend our Master Motor Map (MMM) framework [10] to offer methods and tools needed for processing the data and make the dataset publicly available in our KIT Whole-Body Human Motion Database[1], see also [11].

II. Related Work

We first review related human motion data collections in the close context of multi-modal bimanual recordings of daily household and kitchen activities. The datasets are categorized based on the used sensor modalities. We provide an overview of the most relevant related works in Table I. Our comparison is based on provided action annotations, sensor modalities, especially the accuracy of whole-body pose and object interaction, and captured variations within an action type. Many datasets provide unconstrained recordings of various subjects performing naturally in unstructured environments to capture a wide variance in data across all dimensions. However, the introduction of explicit single variations in object types and relations, as well as bimanual execution, within actions is beneficial for research on generalizing bimanual task models, as it allows better comparison of the influence of different task parameters on the execution.

A. Single-View Video Datasets

Several large-scale datasets with video recordings of humans performing various actions in different daily scenarios are available.

Head-mounted video cameras are often used to record a human subject's field of view and create datasets of daily activities in natural environments (e. g. [12]–[15]) because they are easy to attach to the human body and continuously capture the workspace of the subject even for mobile manipulation. The *EPIC Kitchen Dataset-100* [12] with 100 hours of long-term unscripted kitchen activities is the largest annotated egocentric action dataset. The *20BN-Something-Something dataset* [16] offers a similarly large collection of very short video clips containing first- and third-person human-object interactions. However, both datasets do not provide recordings of the human whole-body motion and have been collected in an unknown or changing camera coordinate system. In [17] the *YouCook* dataset is created by collecting and annotating 88 open-source third-person cooking videos from YouTube. In comparison, the *MPII Cooking Activties Dataset* [18] provides annotated 27 hours of self-recorded static camera videos of subjects preparing real dishes in a kitchen environment.

Overall, these datasets often provide large amounts of videos in various, unstructured environments because the data collection effort is comparatively low. The recordings are either collected from various open video platforms, or recorded in real-world scenarios as neither a large sensor setup nor special equipment is required. Because of their size, such datasets are often used for training and evaluation of machine learning algorithms, especially in the context of action recognition, detection, anticipation and retrieval.

Methods for extracting 2D [19] and 3D [20] human poses, grasp types [14] as well as object bounding boxes [21] and 3D poses [22] of known objects from RGB videos can be used to obtain various information from video data, however, retrieving accurate information in various scenarios under different conditions is still difficult. Further, learning robot manipulation concepts based on video datasets as described in e. g. [23] can benefit from datasets that provide comprehensive knowledge about human demonstrations.

B. Multi-View and/or Multi-Modal Video Datasets

Collecting multi-view video datasets or using additional sensor modalities to capture human demonstrations contribute to the extraction of further knowledge.

The *TUM Kitchen Data Set* [24] is based on the fusion of multiple RGB camera streams to recreate the three-dimensional human pose. In addition, RFID tags and magnetic sensors are used to detect subjects opening a door or a drawer while setting a table. In *Slice & Dice* [25] three-axis accelerometer measurements from sensorized cooking

[1] https://motion-database.humanoids.kit.edu/

TABLE I: Overview of human motion datasets for object manipulation.

	General				Label		Action Variations					Sensors																										
												Marker-based Motion Capture					Vision				Glove			Attached														
																								Type								Location						
Reference	Year	Household actions	Bimanual execution	3D object models	Fine-grained actions	Separate per hand	Unconstrained	Spatial variations	Tracked objects	Bimanual execution	Other[3]	Human whole-body	Hand configuration	Face mimics	Primary object	Secondary objects	RGB	Depth	Multi-view	Egocentric	Pressure	Kinematics	Both hands	Gyroscope	Magnetometer	Accelerometer	Torque sensor	Force sensor	Pose sensor	Other[4]	Human whole-body	Human arm	Other body parts	Objects	Environment	Audio
[34]	08	x	x				x					x					x		x	x				x	x	x						x	x				x
[24]	09	x	x		x	x	x										x		x											x					x		
[25]	09	x	x		x		x																			x							x				
[38]	10	x	x		x	x	x																	x	x	x				x	x		x	x	x		x
[26]	13	x	x		x		x										x	x								x											
[27]	14	x	x	x			x										x	x																			x
[15]	14	x	x				x										x			x																	
[11]	15	x[1]	x[1]	x				x	x		x	x			x	x	x																				
[16]	17	x[1]	x[1]				x										x			x[1]																	
[37]	19		x		x						x	x					x		x		x	x[1]		x	x	x				x							
[39]	19	x						x	x		x				x[2]	x[2]	x[1]	x[1]				x					x	x	x[2]				x				
[31]	20	x	x		x	x	x										x	x																			
[12]	20	x	x		x		x										x			x																	
[30]	20	x	x		x		x										x	x	x	x																	
[41]	20	x[1]	x[1]	x			x					x	x	x	x																							
[29]	20	x	x		x		x										x	x	x																		
Our	21	x	x	x	x	x		x	x	x		x			x	x	x	x	x	x		x	x	x	x						x	x					

[1] only partially [2] switched from motion capture to pose sensor [3] mass, speed, etc. [4] RFID, power, pressure sensor, etc.

utensils are collected during the preparation of sandwiches and salads. Similarly, the 50 *Salads dataset* [26] provides task recordings for preparing various salads, also containing rough object trajectories obtained from accelerometers. In [27] subjects are observed when preparing cereal from multiple view-points with audio signals recorded as an additional sensor modality. In another approach, an egocentric RGB-D camera is used to collect data to classify grasp types and to predict contact points and forces [28].

The *ETRI-Activity3D* dataset [29] focuses on recording motions for recognizing daily activities of elderly people. Therefore, RGB-D videos of a large group of young and elderly subjects are recorded from eight points of view. In the *LEMMA* dataset [30], two static RGB-D and two egocentric RGB cameras are used to record two agents cooperating to perform given tasks in different kitchen and living room environments. While the aforementioned datasets include videos with both hands, the focus is not explicitly on bimanual manipulations. In our previous work we provided the *Bimanual Actions Dataset* [31] that includes only RGB-D videos from a single camera. The focus of this dataset was primarily on human action recognition in bimanual household tasks based on spatial relations. Thereby, action labels were assigned to each hand individually to increase the granularity of the provided annotation.

Similar to larger single video datasets, most of the described datasets offer mostly unconstrained motion recordings for training and evaluation of action recognition and prediction methods. In addition, such datasets are essential for learning from human observation tasks such as e. g. identifying changes in 3D semantic relations during bimanual manipulation [31], learning simple motion primitives [32] or learning of object affordances [33]. However, in complex bimanual manipulation scenarios more suitable motion tracking approaches are needed to deal with multiple small or feature-less objects and occlusions to allow extracting accurate motion trajectories of subjects and objects.

C. Motion Capture Datasets

Several large-scale human motion databases are available [11], [34]–[36] but most are less focused on explicitly capturing object manipulation.

The *Carnegie Mellon University Multimodal Activity* (CMU-MMAC) database [34] consists of recordings obtained with various sensor modalities capturing subjects performing cooking and food preparation tasks. In the *AnDyDataset* [37] industry-like activities such as screwing and manipulating loads under various conditions are recorded with a multimodal sensor setup for the purpose of human motion classification, prediction and evaluation in industrial environments. However, both datasets only provide object motion in the video data. In the *OPPORTUNITY Activity Recognition Data Set* [38], the subjects perform daily life tasks while their pose is tracked with inertial measurement units (IMUs), and interactions with objects and the environment are captured with a variety of sensor modalities.

In comparison, the *Daily Interactive Manipulation (DIM) Dataset* [39] focuses on interactive manipulation, particularly motions where an object or tool is manipulated by the subject to perform an interaction with another object. For this purpose, a custom-built handle including a force-torque sensor is attached to the manipulated object/tool. A large

number of short actions, especially pouring actions and their variations (objects, content) were recorded. In addition to force information, object poses are also captured. While this dataset contains subsymbolic information (position and force trajectories) of various daily manipulation actions, only unimanual motions are performed using the sensorized tool, which also prohibits a natural grasping and manipulation behavior. Further, the dataset does not include human body motion during the task execution. In [40], recordings of 37 subjects performing a "fruit scooping" task are collected. Similar to the previous dataset, a sensorized tool is used, but in addition the corresponding human hand and forearm are tracked with multiple sensors including motion capture. In this case, a bimanual task is considered in which the other hand kinesthetically guides a robotic arm to hold the fruit.

The *GRAB: GRasping Actions with Bodies* dataset [41] provides a large amount of marker-based motion capture recordings containing full 3D human shape and pose sequences including the hand and face motion of a subject interacting with 51 different 3D printed objects. The emphasis is put on whole-body grasping and the estimation of actual grasp contact surfaces. In this dataset, only the interaction with a singular object is considered during bimanual grasping and actions. While offering the opportunity for researching bimanual grasping and handover tasks, the dataset is less suited for analyzing and learning bimanual actions in the context of goal-oriented object-object interaction.

Compared to related work, we provide a new multi-modal dataset of whole-body motions for learning task models of bimanual manipulations. The new dataset complements our Whole-Body Human Motion Database [11] and the Bimanual Action Dataset [31] with recordings of bimanual tasks. Recent work has not only emphasized the role of the whole-body posture in the ability to perform task-specific motions and exert forces but also provides methods for transferring such uni- and bimanual manipulation tasks to robots [42]. The design of our dataset is motivated by the goal of providing all the necessary information for learning such task models from human demonstration with explicit focus on capturing variations in object types and object relations in bimanual tasks. Further, we provide manual action annotations for each hand.

III. The Dataset

This section explains the sensor setup, recorded objects, action types and variations. It also describes the synchronization of this dataset and how experiments were conducted. More detailed information, such as the exact marker setup or precise anthropometric subject data, as well as all performed motion recordings are available in the KIT Whole-Body Human Motion Database [11].

A. Sensor Setup

The sensor setup consists of a total of five different sensor modalities. A marker-based VICON motion capture system is used to capture accurate trajectories of body segments and objects at a frequency of 100 Hz. Figure 2 provides an overview of the camera setup in our motion capture lab, which is equipped with nine static motion capture cameras (*MX T10*) attached to the wall around the capture area at a height of about three meters and mobile cameras (*Vero*), which are placed on tripods around the subject of interest. As shown in Figure 1, the subject wears a full body suit (Prophysics AG and NaturalPoint) with attached optical markers of a diameter of 14 mm that are tracked by the infrared cameras. In parallel, the experiments are recorded with a connected digital video (DV) camera for documentation.

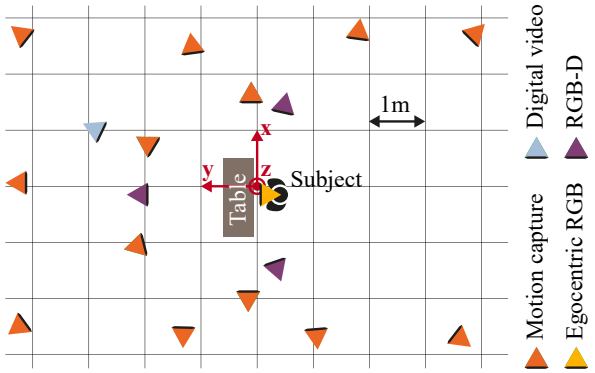


Fig. 2: Positions of employed cameras with respect to the subject and table. The bold line depicts in which direction the camera is facing. The red coordinate system marks the origin of the motion capture system.

For the recording of hand grasping movements in bimanual tasks, the subjects wear commercially available data gloves (Cyber Glove III) on each hand. These data gloves measure the finger joint angles, the palm curvature and the wrist angles. We used the currently available data gloves in our lab, which is a right hand data glove with 22 degrees of freedom (DoF) and a left hand data glove with 18 DoF (all DoFs except the distal finger joints). The data gloves are calibrated as described in our previous work [43] and capture finger joint positions at a frequency of 90 Hz. We consider using such data gloves as the most suitable way to capture finger position trajectories in bimanual manipulation tasks as capturing such trajectories with the same marker-based motion capture system would require a high number of additional markers on both hands and lead to occlusions and wrong marker associations when the hands closely interact with multiple objects.

In addition, three 9 DoF inertial measurement units (IMUs) (Blue Trident of Vicon Motion Systems) are attached to the human body to measure linear accelerations and angular velocities at a frequency of 225 Hz. The data is upsampled to 300 Hz in order to obtain an integer number of sub-samples per frame of the VICON system. The sensors are attached to the human body at anatomical landmarks: one sensor on each forearm close to the wrist (dorsal side of the *antebrachium above carpals*) and one sensor on the back between the shoulder plates (approx. *thoracic vertebrae T4*).

Fig. 3: Objects used in our bimanual manipulation dataset.

To obtain egocentric images from the perspective of the subject, *Full HD* video recordings are collected with a head-mounted action camera (GoPro Hero 8) with 60 FPS in *SuperView* and activated *HyperSmooth* for automatic image stabilization.

In addition, three RGB-D cameras (Azure Kinect DK) are placed at fixed positions on camera tripods. The camera positions are selected to provide different views on the scene and indicate potential positions of a robot that should learn from human demonstration. These video recordings are obtained at 30 FPS, 1080p RGB resolution and 640 px × 576 px depth resolution. To obtain and track the pose of the RGB-D and action cameras in respect to the scene, optical markers are attached to the cameras to record their pose.

B. Objects

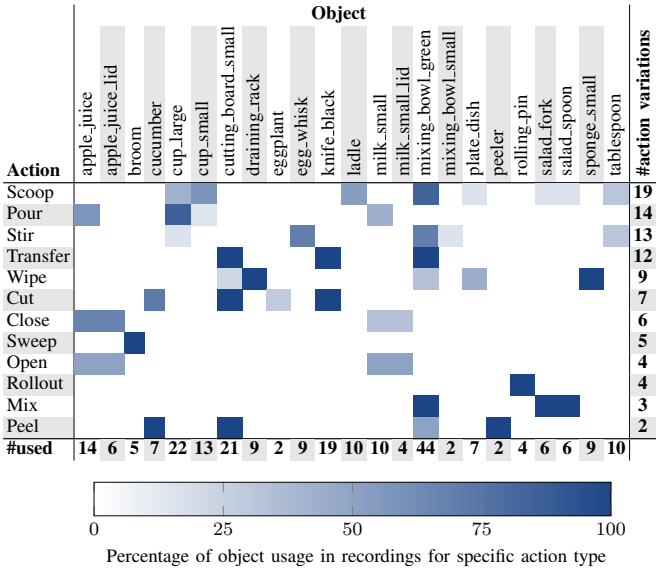
The actions were recorded with a total of 21 real household objects and food items (e. g. cucumber, knead dough), see Figure 3. All actions were performed on or behind a table with a height of 88 cm, which corresponds to the height of common kitchen counters. At least four markers were attached to each object in order to track its pose with the motion capture system. Depending on the object, markers of size 6 mm, 9.5 mm or 14 mm were used. 3D models of all objects are provided with the dataset. These models were created either using a 3D scanner or CAD software in the case of simple object geometry.

C. Actions

Twelve frequently in household activities used manipulation actions were selected for the dataset with focus on the variety of used objects and the bimanuality of actions. The actions, the number of recorded variations as well as used objects per action are given in Table II, where the color code indicates the object usage for an action across all variations.

Although some actions such as pouring can be performed using only a single hand, this dataset explicitly focuses on recording the execution of such tasks with both hands. In many cases these are asymmetric movements [44], in which one hand stabilizes an object while the other hand performs the manipulation of the object. For example, the left hand holds a cup on the table or in midair while the right hand pours water from a bottle. There are also cases where both hands are holding the same object (e. g. sweeping, rolling out dough) or self-handovers are performed. Other special cases

TABLE II: Recorded actions and used objects



are walking while performing a manipulation or holding an object by enclosing it with the whole arm.

All actions are recorded in the way they would be naturally performed by right-handed individuals. To account for variations in natural human action execution, every variation was repeated three times by the subject. Also, semantic variations within one action type are considered, such as different object locations. For example, the bottle can be on the left or right side of the cup and next to or farther away from the subject. Such variations are important to investigate the adaptation of motion trajectories to new scenes and situations. Furthermore, different objects are included focusing especially on changes in single task parameters such as the object height or diameter (e. g. small vs. large cup or bowl). Different actions with tools are also considered such as scooping with a spoon or ladle. The tools are handled differently even though they are used for the same purpose. Additionally, different executions of bimanual manipulation actions are considered. For example, the left hand holds, tilts or lifts a bowl while the right hand stirs. For some actions, the object held in hand during manipulation can also be different. As shown in Table II there is an imbalance in the number of recorded variations per action type. This is due to the fact that for some actions interesting parameters (e. g. position, direction, height) allow more variations in the execution.

D. Data Synchronization

To provide timely-synchronized data from the different sensor modalities used for the collection of the dataset, we implemented a software named *CaptureComponent*, which synchronously triggers the recordings of the different sensors on several remote host computers. The component distributes signals as UDP packages to remote software components of selected sensors and manages recording-specific information such as filenames or recording time. These software components are implemented for each sensor and encapsulate

the sensor-specific processing and functionality to enable independence of the programming language of the sensor's interface, the operating system and hardware. The *CaptureComponent* also provides access to each distributed component to check the sensor status. Further synchronization is achieved by collecting and aligning timestamps for each sensor recording.

E. Recording Procedure

Two healthy, young adults (1 male, 1 female) participated in the experiments. Both are right-handed, have normal vision and no upper limb orthopedic impairments. Anthropometric data of the subjects (body height, hand segment lengths, weight) are recorded. The dataset includes recordings of only two subjects due to the high efforts required for providing such high quality multi-modal data with a large number of action variations and repetitions. Subjects were familiar with the task but were asked to perform the actions as they would execute them in their own home environment.

The subject stands behind the table at the beginning and the end of each recording with flat placed hands on the table. Start and end configuration of the scene, such as hand and body posture, and task instruction, such as *'cut off three slices of the cucumber'*, are given to the subjects. Details of the exact task execution such as the temporal synchronization of the hands and grasp types are left to the subject's intuition. Three repetitions of each action were recorded. In total, we collected 98 demonstrations per subject with three repetitions each, resulting in 588 demonstrations.

Within the recordings of the different variations of an action the order of the recordings was the same among the subjects, but the actions themselves (e. g. *Scoop*) were recorded in a different order. The action durations, i. e. the length of the recordings, range between 5 and 15 seconds. The different actions were recorded on multiple days.

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. For data privacy reasons the faces of subjects are blurred in all publicly available visual recordings.

IV. Representation and Processing of the Data

In order to make the collected data from various sensors available to the research community, we rely on and expand our previous work on the Master Motor Map (MMM) framework [10] to provide a unified representation and standardized data structures for organization and storage in large scale motion databases. In the following, we describe the extensions made to the MMM to address the needs of the new recordings as well as the tools provided for segmentation and labeling of the dataset.

A. MMM Framework

The Master Motor Map (MMM) [10] provides an open-source framework for the representation of human motions as well as their perception, visualization, reproduction and recognition. The MMM framework[2] decouples the motion capture process from further processing steps by providing a reference model of the human body as well as a unifying motion data format. Through the use of a reference kinematics and dynamics model with subject-specific parameters, captured motions are normalized and presented in a standardized way. For the mapping of captured human motions to a target reference embodiment, the squared distances between real markers attached to the subject's body at predefined anatomical landmarks and virtual markers on the MMM reference model at the same anatomical landmarks are minimized using non-linear optimization techniques.

In this work, this mapping is extended by considering the hand size of the subject by scaling the hand model in the MMM independently of the human's height. The hand size is determined by measuring the distance from the subject's wrist to the tip of the middle finger. To further improve the accuracy of the hand pose mapping, the squared error of each hand marker is scaled with a specific weighting parameter. We provide one possible mapping of the motion in this dataset. However, this mapping can be exchanged or adjusted to address the needs of the intended application.

Moreover, the XML-based MMM data format as well as the framework is adapted to independently store, handle and visualize data from all additional sensor modalities in this dataset such as IMU, RGB, RGB-D and data gloves using an extendable plugin-based sensor structure. The videos are stored in a suitable video container format (e. g. mp4) and are only referenced in the MMM data format.

B. Segmentation and Labeling

In order to provide a suitable interface for further processing of the recorded human movement, the MMM data format is extended to allow both manual and automatic hierarchical segmentation, i. e. on symbolic and subsymbolic level as proposed in [45], as well as annotation of the data. Similar to the previous work of [31], the motion recordings in this work are manually segmented and annotated according to actions performed by each hand.

Manipulation tasks are usually composed of several actions such as approach an object, lift it, perform a manipulation, place the object and retreat the hand. The annotations include manipulation actions (e. g. *scoop, wipe, peel*), supporting actions (e. g. *hold, move*) and actions describing different grasping phases (*approach, lift, place, retreat*). An example segmentation for scooping is displayed in Figure 4. The manipulation task is segmented into the following actions for the right hand: *approach ladle with right hand, lift ladle, move hand with ladle in the bowl, scoop, move hand with ladle to the cup, pour from the ladle into the cup, ...* As displayed, all manipulation actions such as *scoop* are further hierarchically segmented. These fine-granular segments can be used when considering isolated partial actions.

Relevant objects are also included in the annotation. We distinguish between a *main* object that corresponds to the

[2] `https://mmm.humanoids.kit.edu`

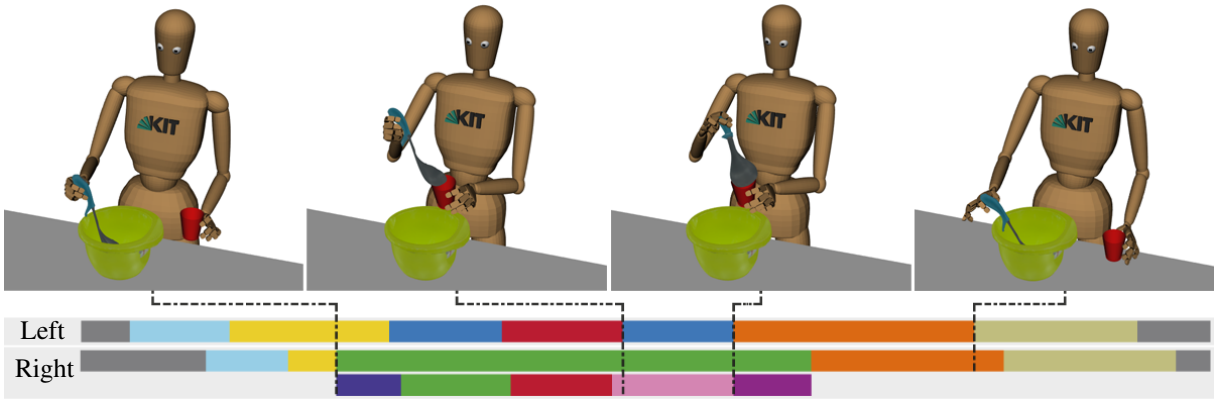


Fig. 4: Bimanual segmentation for scooping from a bowl to a cup. *Top:* Visualisation of the motion on the MMM reference model. *Bottom:* Segmentation tracks for both hands (Left, Right). Occuring actions are idle, approach, lift, hold, move, place, retreat, scoop as well as additional actions pre, pour and post scooping task.

grasped object, a *target* object with which the main object interacts, and a *source* object that is only given if such interaction occurs at the beginning of a manipulation task. In the example of the manipulation action *"scooping from a bowl into a cup using a ladle"*, the bowl corresponds to the source and the cup to the target object, while the ladle is the main object.

V. Conclusion

We present a new multi-modal dataset of bimanual manipulation actions, which has been recorded from human demonstrations using several sensors with focus on providing all information needed for learning task models of bimanual manipulations from human demonstration on symbolic and subsymbolic level. Thus, precise whole-body human motions including both hands as well as object motions are represented in a unifying way using the Master Motor Map (MMM) framework and stored publicly available in the KIT Whole-Body Human Motion Database.

The dataset contains 12 different bimanual actions, performed by two subjects with up to 19 variations per action and three repetitions of each action variation. This results in a total of 588 human demonstrations. The human whole-body pose, hand kinematics as well as 21 household objects and real food items are tracked using marker-based motion capture and data gloves. In addition, data of three inertial measurement units, one head-mounted action camera, and three RGB-D cameras is collected simultaneously.

In particular, the dataset provides intra-action variations representing different ways of human demonstrations to facilitate learning and generalization of bimanual manipulation actions while taking into account multiple modes and models in the demonstrations as these are key for task-specific adaptation and generalization of movement primitives representing the underlying actions in the demonstrations [46]. While focusing on bimanual manipulation tasks, we provide whole-body motion and full hand configurations to facilitate research on whole-body loco-manipulation tasks.

Furthermore, the MMM framework [10] and data format is extended to deal with the synchronization and normalization of the recorded multi-modal data, as well as the mapping of this data to the MMM reference model. In addition, we provide tools within the MMM framework for manual action segmentation and annotation of bimanual manipulation tasks.

The dataset, tools and methods have been collected and developed to facilitate research in the area of learning bimanual task models from human observation and will contribute to many research directions in this area ranging from learning motion primitives, action and activity recognition, learning spatial and temporal constraints in bimanual human demonstrations as well as bimanual coordination.

Our future work will be concerned with the continuous extension of the dataset regarding the number of actions and their variation, as well as the number of recorded subjects performing longer demonstrations. Further, we will work on learning task models for bimanual manipulation using the dataset, in particular on incremental learning of such models, temporal relations and transitions between actions of both hands, and on defining related benchmarks in this research area.

Acknowledgment

The authors would like to acknowledge the help and support of our motion capture students (Kevin Liang, Adnan Ügür, Mitat Hoxha and Saghar Samaei) of the H^{2}T lab who contributed to this 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, "Survey: Robot programming by demonstration," *Handbook of robotics*, vol. 59, 2008.
[2] A. Billard and D. Kragic, "Trends and challenges in robot manipulation," *Science*, vol. 364, no. 6446, 2019.
[3] A. Ude, A. Gams, T. Asfour, and J. Morimoto, "Task-specific generalization of discrete and periodic dynamic movement primitives," *IEEE Transactions on Robotics*, vol. 26, no. 5, pp. 800–815, 2010.
[4] P. Pastor, H. Hoffmann, T. Asfour, and S. Schaal, "Learning and generalization of motor skills by learning from demonstration," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2009, pp. 763–768.
[5] S. Calinon, "Learning from demonstration (programming by demonstration)," *Encyclopedia of Robotics*, pp. 1–8, 2018.

[6] S. Oprea, P. Martinez-Gonzalez, A. Garcia-Garcia, J. A. Castro-Vargas, S. Orts-Escolano, and J. Garcia-Rodriguez, "A visually realistic grasping system for object manipulation and interaction in virtual reality environments," *Computers & Graphics*, vol. 83, pp. 77–86, 2019.

[7] C. Uhde, N. Berberich, K. Ramirez-Amaro, and G. Cheng, "The Robot As Scientist: Using Mental Simulation to Test Causal Hypotheses Extracted from Human Activities in Virtual Reality," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2020, p. 6.

[8] A. Haidu and M. Beetz, "Automated acquisition of structured, semantic models of manipulation activities from human vr demonstration," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[9] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid, "Learning from synthetic humans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 109–117.

[10] 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 Transactions on Robotics*, vol. 32, no. 4, pp. 796–809, 2016.

[11] ——, "The kit whole-body human motion database," in *2015 International Conference on Advanced Robotics (ICAR)*. IEEE, 2015, pp. 329–336.

[12] D. Damen, H. Doughty, G. M. Farinella, A. Furnari, E. Kazakos, J. Ma, D. Moltisanti, J. Munro, T. Perrett, W. Price, *et al.*, "Rescaling egocentric vision," *arXiv preprint arXiv:2006.13256*, 2020.

[13] A. Fathi, X. Ren, and J. M. Rehg, "Learning to recognize objects in egocentric activities," in *CVPR*. IEEE, 2011, pp. 3281–3288.

[14] M. Cai, K. M. Kitani, and Y. Sato, "An ego-vision system for hand grasp analysis," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 4, pp. 524–535, 2017.

[15] I. Bullock, T. Feix, and A. Dollar, "The Yale human grasping dataset: Grasp, object, and task data in household and machine shop environments," *The International Journal of Robotics Research*, vol. 34, pp. 251–255, 2014.

[16] R. Goyal, S. E. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, *et al.*, "The "Something Something" Video Database for Learning and Evaluating Visual Common Sense," in *ICCV*, vol. 1, no. 4, 2017, p. 5.

[17] P. Das, C. Xu, R. F. Doell, and J. J. Corso, "A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2634–2641.

[18] M. Rohrbach, A. Rohrbach, M. Regneri, S. Amin, M. Andriluka, M. Pinkal, and B. Schiele, "Recognizing fine-grained and composite activities using hand-centric features and script data," *International Journal of Computer Vision*, vol. 119, no. 3, pp. 346–373, 2016.

[19] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[20] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt, "Vnect: Real-time 3d human pose estimation with a single rgb camera," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–14, 2017.

[21] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[22] K. Pauwels and D. Kragic, "Simtrack: A simulation-based framework for scalable real-time object pose detection and tracking," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015, pp. 1300–1307.

[23] L. Shao, T. Migimatsu, Q. Zhang, K. Yang, and J. Bohg, "Concept2robot: Learning manipulation concepts from instructions and human demonstrations," in *Proceedings of Robotics: Science and Systems (RSS)*, 2020.

[24] M. Tenorth, J. Bandouch, and M. Beetz, "The TUM kitchen data set of everyday manipulation activities for motion tracking and action recognition," in *IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, 2009, pp. 1089–1096.

[25] C. Pham and P. Olivier, "Slice&dice: Recognizing food preparation activities using embedded accelerometers," in *European Conference on Ambient Intelligence*, 2009, pp. 34–43.

[26] S. Stein and S. J. McKenna, "Combining embedded accelerometers with computer vision for recognizing food preparation activities," in *Proceedings of the ACM international joint conference on Pervasive and ubiquitous computing*, 2013, pp. 729–738.

[27] A. Pieropan, G. Salvi, K. Pauwels, and H. Kjellstrm, "Audio-visual classification and detection of human manipulation actions," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2014, pp. 3045–3052.

[28] G. Rogez, J. S. Supancic, and D. Ramanan, "Understanding Everyday Hands in Action from RGB-D Images," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 3889–3897.

[29] 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 International Conference on Intelligent Robots and Systems (IROS)*, pp. 10 990–10 997, 2020.

[30] B. Jia, Y. Chen, S. Huang, Y. Zhu, and S.-C. Zhu, "LEMMA: A Multi-view Dataset for LEarning Multi-agent Multi-task Activities," in *European Conference on Computer Vision*, 2020, pp. 767–786.

[31] 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.

[32] A. C. Dometios, Y. Zhou, X. S. Papageorgiou, C. S. Tzafestas, and T. Asfour, "Vision-based online adaptation of motion primitives to dynamic surfaces: application to an interactive robotic wiping task," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1410–1417, 2018.

[33] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from rgb-d videos," *The International Journal of Robotics Research*, vol. 32, no. 8, pp. 951–970, 2013.

[34] F. de la Torre, J. Hodgins, A. Bargteil, X. Martin, J. Macey, A. Collado, and P. Beltran, "Guide to the carnegie mellon university multimodal activity (cmu-mmac) database," in *Tech. report CMU-RI-TR-08-22, Robotics Institute, Carnegie Mellon University*, April 2008.

[35] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1325–1339, 2014.

[36] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "AMASS: Archive of Motion Capture as Surface Shapes," in *International Conference on Computer Vision*, 2019, pp. 5442–5451.

[37] P. Maurice, A. Malais, C. Amiot, N. Paris, G.-J. Richard, O. Rochel, and S. Ivaldi, "Human movement and ergonomics: An industry-oriented dataset for collaborative robotics," *The International Journal of Robotics Research*, vol. 38, no. 14, pp. 1529–1537, 2019.

[38] D. Roggen, A. Calatroni, M. Rossi, T. Holleczek, K. Förster, G. Tröster, P. Lukowicz, D. Bannach, G. Pirkl, A. Ferscha, *et al.*, "Collecting complex activity datasets in highly rich networked sensor environments," in *2010 Seventh international conference on networked sensing systems (INSS)*. IEEE, 2010, pp. 233–240.

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

[40] L. P. Ureche and A. Billard, "Constraints extraction from asymmetrical bimanual tasks and their use in coordinated behavior," *Robotics and autonomous systems*, vol. 103, pp. 222–235, 2018.

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

[42] N. Jaquier, L. Rozo, and S. Calinon, "Analysis and transfer of human movement manipulability in industry-like activities," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 11 131–11 138.

[43] J. Starke, C. Eichmann, S. Ottenhaus, and T. Asfour, "Synergy-based, data-driven generation of object-specific grasps for anthropomorphic hands," in *IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)*, 2018, pp. 327–333.

[44] Y. Guiard, "Asymmetric division of labor in human skilled bimanual action: The kinematic chain as a model," *Journal of motor behavior*, vol. 19, no. 4, pp. 486–517, 1987.

[45] M. Wächter and T. Asfour, "Hierarchical Segmentation of Manipulation Actions based on Object Relations and Motion Characteristics," in *International Conference on Advanced Robotics (ICAR)*, 2015, pp. 549–556.

[46] Y. Zhou, J. Gao, and T. Asfour, "Movement primitive learning and generalization using mixture density networks," *IEEE Robotics & Automation Magazine*, vol. 27, no. 2, pp. 22–32, 2020.