Recent Datasets on Object Manipulation: A Survey

Yongqiang Huang, Matteo Bianchi, Minas Liarokapis and Yu Sun

Abstract—**A dataset is crucial not only for model learning and evaluation but also to advance knowledge on human behavior, thus fostering mutual inspiration between neuroscience and robotics. However, choosing the right dataset to use or creating a new dataset is not an easy task, due to the variety of data that can be found in the related literature. The first step to tackle this issue is to achieve a good knowledge of those that are available. In this work, we take a significant step forward by reviewing datasets that were published in the last 10 years and that are directly related to object manipulation and grasping. We report on modalities, activities, and annotations for each individual dataset and we discuss our view on its use for object manipulation. We also compare the datasets and summarize them. Finally, we conclude the survey by providing suggestions and discussing the best practices for the creation of new datasets.**

I. Introduction

Big and organized datasets are valuable in various scientific fields, primarily because they are crucial for revealing hidden patterns, testing hypotheses, and evaluating algorithms. The demand for datasets follows the advancement of a multi-disciplinary field or the evolution of particular problems, and new datasets never stopped being created. In robotics grasping and manipulation, many datasets were recently created by a number of groups for different research purposes and very often shared in the robotics community. It is possible to find human motion datasets [1], instrumental activities of daily living (IADL) datasets [2] for, object and model datasets [3], object geometry and motion datasets [4], haptic interaction datasets [5], among others. The datasets are not only crucial for evaluating and comparing the performances of novel methods [5], but they are also extremely valuable for motion / path planning (see e.g. [6] for a review), robotic learning and training [7] and investigation of human behavior. The goal is to achieve a mutual inspiration between neuroscience and robotics, thus leading to the definition of effective design and control guidelines for artificial systems [8].

The role of datasets should be not only to inform the control and development of robotic devices, but also to verify or deny the correctness and effectiveness of an algorithm or system design, and expose the flaws or exemplify the strength of the algorithm or the design itself. However, in order to properly choose a *good* dataset, one first needs to know what datasets are already available, what they include, and how they differ. Then one can decide on whether any dataset would be useful and which one would best serve the research purpose. One may also decide that none of the datasets suits the purpose, and the reason on which that particular decision is made can be used to improve on the existing datasets and prepare new ones. To help one with choosing the right dataset(s) or deciding on creating new datasets, we contribute a review of datasets that we consider useful for research on object manipulation and grasping. The datasets were created no earlier than 2009 since earlier datasets are usually not supported or accessible.

Object manipulation is the process of changing in a controlled fashion the position and orientation of an object in order to execute a specific task. In contrast to a gross motion such as waving and stretching, an object manipulation motion is a fine motion, and the body parts involved cover a much smaller physical space. In this survey, we mainly focus on datasets that contain object manipulation motions. Gross motions may be present in certain reviewed datasets, but they do not play a dominant role. Under this regard, it is worth to mention another type of datasets, which are specifically designed for whole-body motions or limb motions, and hence are not considered in this review. They are particularly important to humanoid, behavioral science, rehabilitation, neuroscience, and human computer interaction. One example is the KIT Whole-Body Human Motion Database [1] `https://motion-database.humanoids.kit.edu/`. It focuses on human and object motions, which are annotated through motion description tags. It contains not only motion captured data in a raw format (e.g. marker motions of the capture system), but also information on the subject anthropometric measurements, the objects used and the environment along with a data interface to transfer motion to different kinematic and dynamic models of humans and robots. Other large-scale motion databases available to the scientific community are reviewed in [1].

Furthermore, given the importance of grasping as one of the key topics in robotics research and the motivation for the development of effective robotic manipulators [9], we also discuss datasets on human grasping. Indeed, grasping not only facilitates manipulation, but also determines the arm motions in many object manipulations as indicated in [10], [11], and hence can be considered as pre-manipulation. As previously mentioned, human grasp datasets can offer useful insights not only to better understand the human behavior but also to shape the design and control of artificial systems [12]. Under this regard, it is worth to mention the concept of *hand postural synergies* [13], i.e. broadly, goal directed kinematic patterns of covariation observed between the human hand joints. The underlying concept for a general geometrical interpretation of synergies is the dimensionality reduction, i.e. the number of degrees of freedom (DOFs) of the human hand that can be controlled in an independent manner is actually smaller than the physical one [14]. This idea has been successfully

Yongqiang Huang and Yu Sun are with the University of South Florida, Tampa, FL, USA. email: `yongqiang@mail.usf.edu`, `yusun@cse.usf.edu`. Matteo Bianchi is with the Research Center "E. Piaggio", University of Pisa, Pisa, Italy `matteo.bianchi@centropiaggio.unipi.it`. Minas Liarokapis is with the GRAB lab, Yale University, New Haven, CT, USA `minas.liarokapis@yale.edu`. Yu Sun is the corresponding author.

applied in robotics: i) to relax the control effort using less control variables [15] and ii) to reduce the dimensionality of the problem in robotic grasp planning with dexterous hands [16], facilitating an on-line grasp synthesis [17]. At the same time, such dimensionality reduction has inspired the design of under-actuated robotic hands [18], [19].

In this paper, we divide the datasets into three categories and present them separately: those that include mostly cooking activities, in Section II, those that include more general activities of daily living (ADL), in Section III, and the datasets on kinematics of human grasping of real, or imaginary objects, in Section IV. Following a common classification of human hand pose reconstruction systems [20], we decided to organize the review of grasping and manipulation datasets considering pure vision-based acquisitions and wearable-based acquisitions as independent entries, as discussed in Section IV. It is worth noticing that marker-based recordings of grasping and manipulation activities are also described in Section IV, since a certain level of wearability is still required due to the usage of optical markers placed on the dorsum of human hand. All datasets are summarized in Table I that classifies the datasets according to the year that they were published. Moreover, the use of similar colors denotes a series of similar datasets like in ([21], [21]+), ([22], [23], [24]), and ([25], [26]). In Table II, we list the number of instances provided in each dataset. When a dataset contains sequences, we report the number of sequences; otherwise, we report the number of data samples.

TABLE I
PUBLICATION YEAR OF DATASETS (INCLUDING DATASETS OF HANDCORPUS)

Year (20–)	09	10	11	12	13	14	15	16
Datasets	[27]	[28]	[25]	[21]	[29]	[30]	[24]	[31]
	[32]	[33]	[13]	[21]+	[34]	[35]	[36]	[37]
	[38]		[39]	[22]	[40]	[41]	[42]	[43]
	[44]		[45]	[23]	[26]	[46]	[47]	[48]
				[49]	[42]	[42]	[50]	[51]
				[13]		[52]		

TABLE II
NUMBER OF INSTANCES PROVIDED BY EACH DATASET

Cooking			ADL			HandCorpus		
Data	Size	Type	Data	Size	Type	Data	Size	Type
[27]	20	ms	[38]	20	ms	[13]	286	rg, g
[32]	218	ms	[44]	150	v	[31]	8739	rg
[45]	28	v	[28]	24	ms	[33]	19	rg, g
[21]	17	ms	[25]	60	ms	[36]	825	rg, g
[21]+	30	ms	[26]	120	ms	[13]	285	g
[22]	44	v	[49]	20	v	[39]	114	g
[23]	256	v	[41]	979	ms	[42]	3694	rg
[24]	273	v	[46]	18,210	g	[42]	1	k
[29]	50	ms	[43]	∼59,000	ms	[42]	300	g
[34]	35	ms	[37]	∼650,000	g			
[40]	88	v	[48]	>1,000	ms			
[30]	67	ms	[47]	∼12,000	g			
[35]	77 hr	ms	[50]	13	ms			
			[51]	193	ms			
			[52]	4	ms			

Meaning of abbreviations in column “Type”: “ms”–multimodal sequence, “v”–RGB video, “g”–grasp, “rg”–reach and grasp, “k”–kinematic model. The type “ms” or “multimodal sequence” refers to sequences that contain multiple modalities.

In the first two categories, we present the datasets in ascending chronological order. For each dataset, we report on the modalities, the activities performed, and annotations, and we give our view on how each dataset relates to object manipulation. After reporting on the datasets one-by-one, we summarize them on the availability of modalities, object identifiability in annotated activities, and the forms of temporal segmentation of annotated activities. We also provide the lists of shared annotated activities for the ADL and cooking datasets, respectively.

The datasets reviewed in category three are hosted in the HandCorpus initiative website [53], an open access repository for sharing data about human and robot hands, with the objective of advancing the state of the art of the analysis of both the biological and the artificial side. The HandCorpus goal, under an engineering point of view, is to devise design guidelines from biology observations for the development of effective robotic devices (see e.g. [19]) and grasp planning algorithms [54]. Among all the datasets in HandCorpus, we select and review nine datasets that have collections of human hand kinematics recorded in grasping and manipulation tasks.

For those who want to further examine the datasets covered in this work, we provide the links to all datasets in Table III.

II. DATASETS OF COOKING ACTIVITY

In this section, we present thirteen datasets of cooking activities. The interest in studying cooking activities is motivated by the large number of interactions with the objects and the external environment that human hands and body usually undergo. The datasets include common visual-based acquisition modalities such as RGB vision and depth vision, as well as modalities that are less common such as skin temperature and body heat. RGB vision is used by all datasets. We first present each dataset individually, describing the different characteristics; data type and size, modalities, equipment, annotations etc. Then, we compare the datasets on their different descriptive fields and discuss their suitability and applicability for Learning from Demonstration (LfD) [55], also known as Programming by Demonstration, or Imitation Learning.

A. *Slice&Dice*

Slice&Dice [27] features four instrumented utensils which include three knives of different sizes and a spoon. Each utensil embeds in its handle a 3-axis accelerometer. Twenty subjects participated and each subject prepared a salad or a sandwich freely using the ingredients provided by the experimenter. The acceleration data are accompanied by RGB videos. We consider embedding accelerometers inside objects a merit as, unlike vision based sensors, they provide acceleration data that belong to a certain object alone, and is readily usable without running object recognition first.

B. *CMU-MMAC*

The CMU-MMAC dataset [32] contains multi-modal cooking activities of five recipies: brownie, eggs, pizza, salad, and

TABLE III
LINKS TO DATASETS

[27]	http://openlab.ncl.ac.uk/publicweb/publicweb/AmbientKitchen/KitchenData/Slice&Dice_dataset/
[32]	http://kitchen.cs.cmu.edu/
[45]	images: http://ai.stanford.edu/~alireza/GTEA/ and the rest: https://www.dropbox.com/sh/q4s6nocyhpnauic/AAAvCTfVPCo1u0vTCOsHGwA_a?dl=0
[21](+)	http://ai.stanford.edu/~alireza/GTEA_Gaze_Website/
[22]	https://www.mpi-inf.mpg.de/departments/computer-vision-and-multimodal-computing/research/human-activity-recognition/mpii-cooking-activities-dataset/
[23]	https://www.mpi-inf.mpg.de/departments/computer-vision-and-multimodal-computing/research/human-activity-recognition/mpii-cooking-composite-activities/
[24]	https://www.mpi-inf.mpg.de/departments/computer-vision-and-multimodal-computing/research/human-activity-recognition/mpii-cooking-2-dataset/
[29]	http://cvip.computing.dundee.ac.uk/datasets/foodpreparation/50salads/
[34]	http://www.murase.m.is.nagoya-u.ac.jp/KSCGR/
[40]	http://web.eecs.umich.edu/~jjcorso/r/youcook/
[30]	http://robocoffee.org/datasets/
[35]	http://serre-lab.clps.brown.edu/resource/breakfast-actions-dataset/
[38]	https://ias.in.tum.de/software/kitchen-activity-data
[44]	http://www.cs.rochester.edu/~rmessing/uradl/
[28]	UCI repository: https://archive.ics.uci.edu/ml/datasets/OPPORTUNITY+Activity+Recognition#, Challenge: http://www.opportunity-project.eu/challengeDataset
[25][26]	http://pr.cs.cornell.edu/humanactivities/data.php
[49]	http://www.csee.umbc.edu/~hpirsiav/papers/ADLdataset/
[41]	https://archive.ics.uci.edu/ml/datasets/Dataset+for+ADL+Recognition+with+Wrist-worn+Accelerometer#
[46]	http://www.eng.yale.edu/grablab/humangrasping/
[52]	http://wildhog.ics.uci.edu:9090/#EGOCENTRIC0%20Intel/Creative
[13]	http://www.handcorpus.org/?p=97
[31]	http://www.handcorpus.org/?p=1596
[33]	http://www.handcorpus.org/?p=100
[36]	http://www.handcorpus.org/?p=1507
[13]	http://www.handcorpus.org/?p=91
[39]	http://www.handcorpus.org/?p=103
[42]	http://www.handcorpus.org/?p=1156
[42]	http://www.handcorpus.org/?p=1298
[42]	http://www.handcorpus.org/?p=1578
[43]	https://sites.google.com/site/brainrobotdata/home/push-dataset
[37]	https://sites.google.com/site/brainrobotdata/home/grasping-dataset
[48]	http://rpal.cse.usf.edu/imd/
[51]	https://github.com/jrl-umi3218/ManipulationKinodynamics
[47]	http://www.gregrogez.net/research/egovision4health/gun-71/
[50]	http://www.hci.iis.u-tokyo.ac.jp/~cai-mj/utgrasp_dataset.html

[21](+) refers to both Gaze and Gaze+

sandwich. The modalities include RGB videos from static and wearable cameras, multi-channel audios, motion capture, inertial measurement units (IMU), RFID, etc. We are not positive on the number of subjects that were involved, but we infer that it is between thirty-nine and forty-five. Each subject prepared all the recipes. The dataset also specifically recorded anomalous accidental events that occurred while cooking. Certain modalities are incomplete for certain recipes performed by certain subjects. Annotations exist for sixteen subjects while preparing brownies and correspond to the videos captured by the wearable camera. The annotations apply the structure of "verb+objectOne+preposition+objectTwo", whose components are assembled using grammar.

Except RFID tagging which merely reports the involvement of certain objects, all modalities are on human, which is contrary to the Slice&Dice dataset [27]. The dataset is rich in data of upper arm motions because of the combined use of motion capture and IMUs, and therefore is suitable for 3D manipulation motion analysis.

C. GTEA

The GTEA dataset [45] includes egocentric videos of four subjects performing seven food/beverage preparing activities. The videos amount to 31,222 RGB images. Annotations consist of simple verbs (such as put, take, pour, etc.) and names of objects (cup, sugar, etc.). Object recognition or manually drawn bounding boxes on objects is required prior to analysis of the object motion.

D. Gaze and Gaze+

The Gaze dataset [21] contains RGB egocentric videos of fourteen subjects preparing meals using provided ingredients on a table. The videos were captured using an eye-tracking camera and therefore are accompanied by gaze data. The Gaze+ dataset [21] (later referred to as [21]+) is an upgrade to Gaze, and provides the two modalities in Gaze plus audio. The videos have higher resolution than Gaze, and were captured in an instrumented kitchen instead of on a simple table. Ten subjects were involved and each one of them prepared a set of seven dishes. Actions and objects were annotated in the same way as in Gaze. Compared to static images, egocentric images have much larger proportions of the image showing object manipulation specifically and contain more detail, which we consider a merit. Analyzing object motion, however, would assume that object tracking has been done.

E. MPII Cooking, Cooking Composite, and Cooking 2

MPII sequentially created three datasets related to cooking: the MPII Cooking dataset [22] which focuses on fine grained activity, the MPII Cooking Compositite dataset [23] which focuses on composite activities composed of basic-level activities, and the MPII Cooking 2 dataset [24] which unifies and is an upgrade of both [22] and [23].

The MPII Cooking dataset involved twelve subjects each preparing one to six out of fourteen dishes, and contains forty-four RGB high-definition (HD) videos with a total length of over eight hours or 881,755 frames. The annotations include sixty-five activities, and 5,609 instances were identified.

The MPII Cooking Composite dataset included all the videos from the MPII Cooking dataset and added 212 newly-recorded videos. Eighteen more subjects than in the MPII Cooking dataset participated. Different from the MPII Cooking dataset, the MPII Cooking Composite dataset annotations include four categories: activities (e.g. verbs), ingredients, tools, and containers, which combined are referred to as "attributes". There exist 218 attributes in the dataset, among which seventy-eight are activities. A total of 49,258 attribute instances have been identified which belong to 12,642 annotated temporal segments.

As a refined superset of [22] and [23], the MPII Cooking 2 dataset contains 273 videos involving thirty subjects. The dataset contains fifty-nine dishes, which consist of fourteen diverse and complex dishes from [22], and forty-five shorter and simpler composite dishes from [23]. A total of 222 attributes exist, among which eighty-seven are activities. 54,774 attribute instances have been identified which belong to 14,105 temporal segments. For the above MPII datasets, the subjects were only told which dish to prepare, which lead to natural activities with much variability.

Of all the datasets we include in this work, the MPII datasets altogether have the largest number of HD videos and annotation instances. Objects and fine actions are annotated in great detail, and 2D poses of upper body are also provided. For vision-based 2D object manipulation analysis, the amount of data and action variability of the MPII datasets can only be rivaled by the Brown breakfast dataset [35], if not unmatched.

F. 50 Salad

The 50 Salad dataset [29] extends Slice&Dice [27] by using accelerometers on more utensils and by including depth videos in addition to RGB ones. Twenty-five subjects participated and each prepared a mixed salad twice, and in each run followed a specific sequence of tasks. The sequences were produced by a statistical activity diagram, which would theoretically enable the same number of samples for each task sequence.

The annotation includes three high-level activities: prepare dressing, cut and mix ingredients, and serve salad. Each high-level activity summarizes several low-level activities, and each low-level activity has -pre, -core, and -post phases, which were annotated respectively. 50 Salad inherits the merit of Slice&Dice [27], involves more subjects, enables 3D analysis with depth videos, and has finer annotations. In that regard, we recommend 50 Salad over Slice&Dice.

G. Actions for Cooking Eggs (ACE)

The ACE dataset [34] contains RGB-D videos of cooking activities for five egg menus, all of which were cooked by each of seven subjects. The labels contain only verbs: break, mix, bake, turn, cut, boil, season, and peel. We include this dataset because it provides fine object manipulation motion, but since objects are not identified in any way, using the dataset would rely on human and object tracking more heavily than other datasets.

H. YouCook

The YouCook dataset [40] consists of eighty-eight RGB cooking videos downloaded from Youtube. All the videos have a third person point of view. Although only seven actions labels are used, as many as forty-eight object labels spanning seven object categories exist, and object tracks are provided. We consider the richness of object labels and the availability of the objects tracks as the merits of the dataset, of which the latter facilitates analysis of fine motion in 2D.

I. Actions for Making Cereal

In [30] the data of eight subjects are included while preparing cereal. The dataset includes multiple modalities, including RGB-D videos, audios, estimated six degree-of-freedom (DOF) object pose trajectories, and object mesh models. We consider the object pose trajectories as the merit of the dataset. No other datasets that we include provide such modality, and using the trajectories alone suffices to conduct analysis on 3D object manipulation.

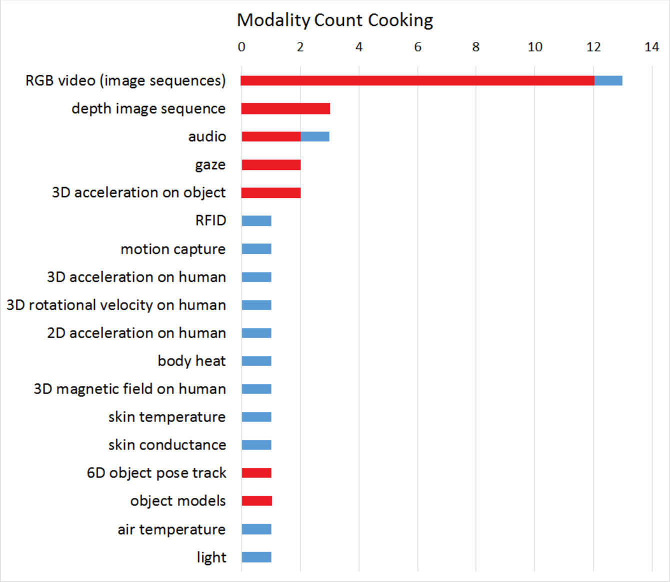

Fig. 1. Count of datasets for each modality. Blue denotes the modalities of CMU-MMAC [32], Red denotes the modalities of the other datasets.

J. Brown Breakfast

The Brown breakfast dataset [35] contains roughly seventy-seven hours of RGB videos involving fifty-two subjects captured at up to eighteen distinct kitchens. In total ten recipes were performed and each subject was reported to have performed all ten recipes, but available data for different subjects vary. Forty-eight coarse activity annotations exist and 11,267 annotation instances were identified. The statistics of the dataset makes it a possible rival of the MPII datasets. It has the largest number of video frames (non HD) among the datasets we include, more than the MPII datasets by 50%. The number of coarse annotation instances is not much lower than the MPII datasets, but the detail and richness of the annotation could not compete with MPII. The dataset does include fine activity annotations, but the statistics and the description of the formation of such annotations are not yet available. Compared with MPII, the dataset lacks 2D upper body pose annotations.

K. Summary

Table IV lays out the different modalities included in all the datasets in this category, and Fig. 1 shows in descending order the count of datasets for each modality.

One can easily notice in Table IV that [32] includes the highest number of acquisition modalities, most of which cannot be found in the other datasets. This is because the goal of [32] is to make the dataset multi-modal.

In Table IV, we can notice that RGB vision is used in all thirteen datasets and is the sole acquisition modality of five

TABLE IV
MODALITIES COOKING

Modalities	[27]	[32]	[45]	[21]	[21]+	[22]	[23]	[24]	[29]	[34]	[40]	[30]	[35]
RGB video (image sequences)	■	■	■	■	■	■	■	■	■	■	■	■	■
depth image sequence									■	■		■	
audio		■			■							■	
RFID		■											
motion capture		■											
3D acceleration on human		■											
3D rotational velocity on human		■											
2D acceleration on human		■											
body heat		■											
3D magnetic field on human		■											
skin temperature		■											
skin conductance		■											
gaze				■	■								
3D acceleration on object	■								■				
6D object pose track												■	
object models												■	
air temperature		■											
light		■											
moving camera		■	■	■	■						■		
# subjects	20	>39	4	14	10	12	30	30	25	7	n/a	8	52

datasets. The equipment required for recording RGB images is generally minimal and is easy to set up. Apart from evaluating certain vision-based algorithms, the recorded RGB video can also be used to verify that the data collection scene is properly set up, to spot any mistakes during the data collection process, and to segment the collected data. RGB images are matrices and carry much more information than other scalar modalities (such as acceleration) captured at a comparable frequency. The pose of the object or the human estimated from RGB images has a lower accuracy than if it is directly measured by a motion capture system, but it usually suffices for action recognition.

Only three datasets include depth images, only two datasets provide information on 3D acceleration on object, and only one dataset provides sequences of estimated object poses. Despite the high accuracy it provides, a motion capture system is used only in [32], possibly because of its cost, the lack of portability and the effort required for the system setup. It is worth to mention that one of the most envisioned applications of these datasets in robotics is for learning from demonstration (LfD). What is commonly done is to use movement sequences of objects as input for training, while testing is performed through physical manipulation of real objects by the robot, (e.g., see [55]). The particular class of applications motivated us to consider and evaluate which characteristics are important to create a dataset that is designed for LfD. Ideally, we would like to have readily available sequences of 3D object poses, which include positions and orientations as in [30]. Object 3D acceleration data can be also converted to 3D position as in [27], [29], and RGB-D images can be used to estimate 3D object pose as in [34]. Another important aspect to take into account in data collection for LfD is the environment, either real or lab-based. For example, among the datasets described in this paper, [40], [21]+, and [35] were collected in real kitchens, and the other datasets were collected either on a table-top or in a lab kitchen. An important difference between a real and a lab kitchen lies in the amount of clutter in the background: real kitchens generally have more clutter, which increases the difficulty of object recognition, which may lower the accuracy of object recognition. Since object pose estimations are fed into LfD as input, a possibly lower accuracy of object recognition is undesirable.

Activity annotations can be useful for various purposes. For example, if the annotations are short sentences describing a video, natural language processing can be combined with vision to provide higher accuracy on action/object recognition, or generate more annotations [56]. Annotations in the form of words can be used to represent motion activity classes. Paulius et al. [57] labeled 65 cooking videos including 798 labeled motion instances and 1229 labeled objects. The knowledge in the cooking videos are represented by a network of motions and objects, which is called functional object-oriented network (FOON). The FOON is continuously updated and maintained at `http://www.foonets.com`. We identified the annotated activities that are shared by multiple cooking datasets, and list those datasets in Table VIII. We combine similar annotations and specify each in the cells.

III. DATASETS OF ACTIVITIES OF DAILY LIVING (ADL)

In this section, we present ten datasets of activities of daily living (ADL), three datasets of grasps acquired using camera, and two datasets of robot motion. The interest of studying ADLs is motivated by the extensive variety of the objects that human hands interacted with daily, and the variety of the environments where these interactions take place. Compared with Section II, this section introduces additional modalities such as 3D kinematics of objects, force and torque on objects and on joints of robotic arms, sequences of estimated human skeleton etc. Apart from action recognition, the application fields of the datasets include hand pose recognition for Human Machine Interaction, grasp analysis, and deep learning, among others. Following the format in Section II, we first review each dataset individually, and then we discuss the use of motion capture and we provide more details on dataset suitability for LfD.

A. TUM Kitchen

The TUM Kitchen dataset [38] contains multi-modal data of set-a-table activities. The modalities include RGB and raw Bayer pattern videos, motion capture, RFID, and reed sensor. Four subjects each transported certain objects from the cupboard, the counter, and the drawer, to a table, and then laid them out in a specified way. The subjects transported the objects one by one as a robot would do, and also several objects at a time as naturally done by a human. The dataset also includes repetitive activities of picking up and putting down objects. The annotations cover the entire duration of the set-a-table activity which starts with *Reaching* through *ReleaseGraspOfSomething*. The actions of the left hand, the right hand, and the trunk were annotated respectively.

Similarly to CMU-MMAC [32], the dataset identifies the objects involved during motion execution, and the availability of motion capture makes it a good candidate for 3D analysis on pick-and-place motion.

B. Rochester ADL

The Rochester ADL dataset [44] contains RGB videos of five subjects performing certain ADL and Instrumented ADL (IADL) activities which can be summarized as: using phone, writing, drinking and eating, and preparing food. Each video records one activity. Similar to the MPII datasets [22]-[24] and the Brown breakfast dataset [35], the Rochester ADL dataset would rely on human and object recognition to be useful for 2D fine motion analysis.

C. OPPORTUNITY

The OPPORTUNITY dataset [28] contains multi-modal data of five morning ADL runs and one Drill run for each of four subjects. Motion sensors were densely deployed on the human body, on the objects, and in the environment. The modalities on the human body include IMUs, 3D accelerometers, and 3D localizers. The modalities on the objects include 3D accelerometers and 2D rotational velocity sensors. The annotations consists of five "tracks": locomotion, high-level activities, mid-level gestures, low-level actions, and objects for the left and the right hand, respectively.

The dataset distinguishes itself from others that we include by using accelerometers and rotational velocity sensors on *both* the hand and the objects. Since object manipulation analysis focuses on the interaction between hand and objects, data that include the motion of both the hand and the objects are desired. The dataset is comparable with 50 Salad [29], CMU-MMAC [32], and TUM Kitchen [38] in modality availability, although the last three target cooking scenarios. For the objects, the dataset includes 2D rotational velocity, which is unavailable in 50 Salad. For the human body, the dataset lacks motion capture, which is available in CMU-MMAC and TUM Kitchen, but alternatively provides 3D acceleration and 3D rotational velocity.

D. Cornell CAD-60 and CAD-120

The CAD-60 [25] and the CAD-120 [26] are both RGB-D video datasets. CAD-60 includes video sequences of four subjects performing twelve ADLs in five different indoor environments. Each sequence corresponds to one instance of a certain activity. The CAD-120 dataset recorded four subjects each performing ten high-level activities. Each subject performed every high-level activity multiple times with different objects. The annotations include ten low-level activities, and twelve object affordances.

CAD-60 and CAD-120 feature skeleton data, which include tracks of 3D position of all fifteen joints plus 3D orientation of eleven joints. The skeleton data in these datasets were generated using the NITE library that complements the PrimeSense sensors and were therefore estimated data. By comparison, the skeleton data collected using a motion capture system are actual physical measurements and therefore can be regarded as ground truth. Thus, the accuracy of the skeleton data in CAD-60 & 120 is lower than the accuracy of those collected with a motion capture system. Nevertheless, the skeleton data are directly usable for 3D fine motion analysis, a characteristic we consider as an advantage of these datasets.

E. First Person ADL

The First Person ADL dataset by Pirsiavash [49] contains RGB videos captured using a GoPro camera. It recorded twenty subjects performing eighteen ADLs. Forty-two objects were annotated by annotators with bounding boxes, tracks, and the status as to whether the object is being interacted with. Similar to Gaze(+) [21], with first person images, the working area of the hands is emphasized. However, since the dataset includes a single modality, using it for analysis on 2D fine motion would rely on object tracking.

F. Wrist-Worn Accelerometer

The wrist-worn accelerometer dataset [41] contains accelerometer data of sixteen subjects performing a total of fourteen ADLs. The accelerometers were attached to the right wrists of the subjects and the data were recorded at the subjects' home. The dataset contains 979 trials. For fine motion analysis, wrist acceleration may be less ideal than hand acceleration, but it remains a readily usable modality.

G. UCI-EGO

The UCI-EGO or general-HANDS dataset [52] includes four sequences of object manipulation activities. Each sequence includes 1,000 RGB-D frames captured using an egocentric camera. Various objects were involved and manipulated, but since the dataset focuses on hand detection and pose estimation, the manipulation tasks performed with each object are relatively short. As other vision oriented datasets, the use of UCI-EGO dataset for object manipulation analysis relies on object tracking.

H. Yale Human Grasping

The Yale human grasping dataset [46] contains 27.7 hours of RGB wide-angle videos of profession-related manipulation motion. Two machinists and two housekeepers participated. The dataset is intended for grasping analysis. The annotations were done on two levels. On the first level, the grasp type was annotated along with the corresponding task name and object name. The second level provided the properties of the object and the task. A total of 18,210 grasp instances have been annotated. The dataset includes prolonged videos of manipulation motion of machining and housekeeping alone, two categories that are not to be found in other datasets that we include.

I. UT Grasp

The UT Grasp Dataset [50] contains data of four subjects who were asked to grasp a set of objects in controlled environment (placed on a desktop) after a brief demonstration of how to perform each type of grasps. A subset of 17 grasp types from Feix's tanonomy were selected which are commonly used in everyday activities [58]. The videos were recorded using a HD head mounted camera (GoPro Hero2) at 30 fps while subjects performed each grasp type with varying hand poses. Annotations were also provided. UT Grasp differs from [46] and [47] in that it consists of data captured in a controlled environment (in front of a desk) in contrast with the [46] and [47] for which data were collected in different parts of a house.

J. GUN-71

The GUN-71 Dataset [47] contains roughly 12,000 RGB-D images of grasps, each annotated with one of the 71 grasp classes in [59]. The images were captured using a chest-mounted camera. 28 objects per grasp were recorded, resulting in 1,988 different hand-object configurations. An important difference between GUN-71 [47] and [46], [50] is that, in [46] and [50] images were captured during daily activities and the annotations follow the distribution of everyday object manipulations, i.e. common grasp classes are much more represented than rare grasps. In contrast, care was taken in GUN-71 to ensure a balanced distribution of grasps and variability of data. To that end, 3-4 different objects were used for each grasp class, 5-6 views of the manipulation scene were considered for each hand-object configuration, 8 subjects participated in the experiments (4 males and 4 females) and 5 different houses were involved.

K. Google Push and Grasping

To facilitate deep learning in robotics, Google Brain publicly shares two datasets of movements of robotic arms: Push [43] and Grasping [37].

The Push dataset contains about 59,000 sequences of multimodal data of robotic arms pushing objects. A bin which contained different objects was placed in front of a 7 DOF robotic arm, and the arm repeatedly pushed the objects in one out of two ways: either pushing randomly, or starting randomly from somewhere on the border of the bin and sweeping towards the middle. A camera was mounted behind the arm facing the bin. The bin contained ten to twenty objects at a time, and the objects were swapped out for new ones after roughly 4,000 pushes. Ten robotic arms were used. The data include RGB images, recorded gripper pose (x, y, z, yaw, pitch), commanded gripper pose, robot joint position and external torques. The dataset provides two test sets each including 1,500 sequences. One test set contains two different subsets of objects from the training set, and the other test set includes two sets of objects absent from the training set.

The Grasping dataset is collected using a similar setup to that of Push. The dataset contains about 650,000 sequences of multimodal data of robotic arms grasping objects. The modalities include RGB-D images, recorded and commanded gripper pose (position in x, y, z and orientation in quaternions), joint positions – velocities – external torques – and commanded torques.

Using Push or Grasping which involve robots only, one aims at learning to finish a task rather than learning to finish a task like a human. The absence of the retargeting problem [60] is an inherent convenience if the learned motion is to be executed by the same robot.

L. Manipulation Kinodynamics

The manipulation kinodynamics dataset [51] includes 3.2 hours of kinematics and dynamics information of objects grasped and manipulated by humans using five fingers. More specifically, the data of the object include mass, inertia, linear and angular acceleration, angular velocity, and orientation. For each of the five fingers, the collected data include friction, force, contact point position, and the axes of a right-handed local coordinate frame $(\mathbf{x}, \mathbf{y}, \mathbf{z})$, where axes $\mathbf{x}$ and $\mathbf{y}$ define the contact surface, and axis $\mathbf{z}$ points towards the object. The dataset does not include images or videos. The objects are custom made and can vary in mass distribution, friction, and shape. The performed motions vary in speed, direction, and task (e.g., emulating pouring). In total 193 different combinations were recorded.

[51] provides a full suite of kinematics and dynamics data. It was created for investigating the mapping relationship between the kinematics features (velocity, acceleration, etc.) of a manipulated object and the underlying manipulating force, which is something similar to a Newtonian physical law. Both the cause of manipulation (the force) and the corresponding result (the kinematics) were measured and both were *of the object*, and no extra processing or estimation is needed. Therefore, we consider the dataset as invaluable for manipulation research, although including RGB-D images would have made the dataset more approachable to the computer vision community.

M. RPAL Tool Manipulation

The dataset [48] features tool manipulation by human and is still in the process of being created. The dataset contains multimodal sequential data of subjects using different tools. The tool consists of four components from front to back: a swappable tooltip, a 6 DOF force-and-torque (FT) sensor, a universal handle, and a 6 DOF position-and-orientation (PO, x, y, z, yaw, pitch roll) tracker. When possible, another PO tracker is mounted on the object which interacts with the tool. Modalities recorded besides FT and PO data are top view RGB videos and depth sequences of the scene, and finger flexture. Currently available data are hosted at `http://rpal.cse.usf.edu/imd/`. Since FT and PO data are of the tool, they can be used directly for manipulation learning, without the need of feature extraction which is necessary for images.

N. Summary

Similar to what we do for the cooking datasets, here we lay out the modalities in all the datasets in Table V, and we show in Fig. 2 the count of datasets for each modality in descending order.

We can see from the figure that, RGB vision is the most commonly used modality and is provided in twelve datasets excluding only [28], [41] and [51]. In fact, [28] did collect RGB videos but did not publish them. Motion capture data are very accurate and can be found in [38] which uses a markerless system, and in [48] which uses both an optical marker-based system and an electromagnetic system. When an object is being manipulated, its orientation may change significantly (for example, when a spatula is used to flip a bread), challenging the reliability of an optical marker-based system. Moreover, objects vary in shape and can be small, which limits the maximum amount of markers that can be used. Also, during the execution of a task, a manipulated object generally has certain contact with another object or certain material (such as water), which makes the contact surface unavailable for mounting markers and the available mounting surface even smaller. The above reasons drove [48] to switch from an optical marker-based system to an electromagnetic (EM) alternative. The EM motion capture system consists of at least one source which defines the world frame and acts as the origin, and one tracker which senses its position and orientation with respect to the source. The source and tracker are both connected to a processing station with cables. Since the EM system uses cables, it does not require an unconstrained line-of-sight as in an optical marker-based system, and therefore a significant object pose change cause occlusions and does not affect measurement accuracy. However, continuously rotating motion such as using a screwdriver has a possibility of finally putting stress on the cable, and therefore requires extra attention.

In this section, we also include RGB vision based grasp datasets [50], [46], and [47], that are quite different from the wearable sensors based datasets presented in Section IV. Vision based algorithms and methods can be used in order to perform both hand pose estimation and classification using simple images of executed grasps or postures. However, the limited estimation accuracy and the lack of directly measured joint angles or fingertips positions limit the applicability of vision based methods in the experimental analysis, design optimization and control of actual robot artifacts (e.g., robot grippers and hands). The vision-based methods are also known to be prone to errors caused by changes in the environmental conditions (e.g., lighting conditions) and their efficiency can also be significantly affected by possible occlusions during the data collection process.

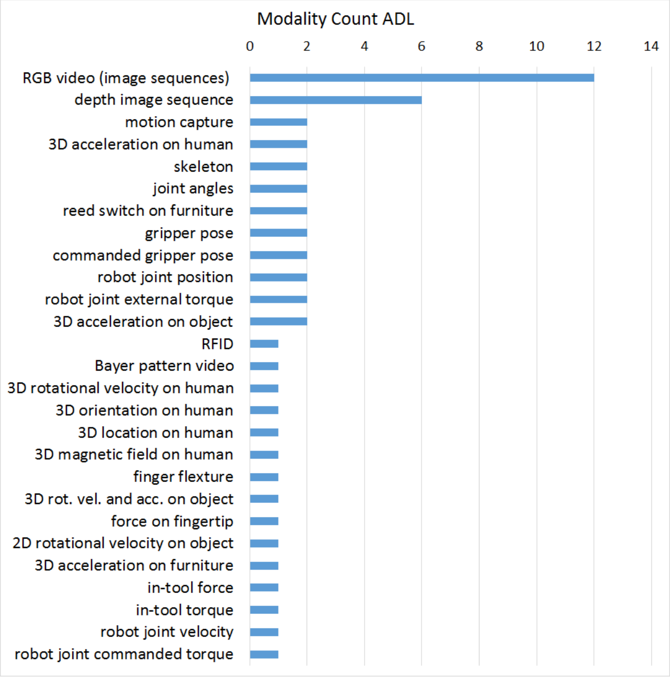

Fig. 2. Count of datasets for each modality

Unique to this section, [51], [48], [43], and [37] introduced the provision of force and torque. The [51] and [48] data belong to the object and the [43] and [37] data belong to the joints of the robotic arms. Including force and torque enables modeling feedback, which makes the learning of object manipulation more physically realistic and helps with performing a learned task with a real object.

The dataset described in [51], is intended for learning the relationship between kinematic features and manipulating force during a manipulation motion *in general*, and not for a particular manipulation task. In simpler words, the dataset focuses on *manipulation* rather than *task*. As a consequence, the dataset falls short of the requirement for Learning from Demonstration [55], which focuses on manipulation *tasks*. Since [51] used 3D printed objects, modifying [51] to make it suitable for LfD would require to change the current 3D object models to enable interaction with other objects while keeping the kinodynamics sensors from interfering with the manipulation tasks, task that may be non-trivial. In comparison, [48] focuses on recording data of tasks and is suitable for LfD, although it provides less fine-grained dynamics data than [51].

As for the cooking datasets, we identified the annotated activities that are shared by multiple ADL datasets, and we list those datasets in Table VI. We combine similar annotations and specify each in the cells. For example, on the first row of Table VI, the annotated activity is summarized as "use phone",

TABLE V
Modalities ADL

Modalities	[38]	[44]	[28]	[25]	[26]	[49]	[41]	[46]	[43]	[37]	[48]	[47]	[50]	[51]	[52]
RGB video (image sequences)	■	■		■	■	■		■	■	■	■	■	■		■
depth image sequence				■	■					■	■	■			■
RFID		■													
motion capture	■										■				
Bayer pattern video	■														
3D acceleration on human			■				■								
3D rotational velocity on human			■												
3D orientation on human			■												
3D location on human			■												
3D magnetic field on human			■												
finger flexture											■				
skeleton				■	■										
joint angles				■	■										
3D acceleration on object			■											■	
3D rotational velocity and acceleration on object														■	
2D rotational velocity on object			■												
3D acceleration on furniture			■												
reed switch on furniture	■		■												
force on fingertip														■	
in-tool force											■				
in-tool torque											■				
gripper pose									■	■					
commanded gripper pose									■	■					
robot joint position									■	■					
robot joint velocity										■					
robot joint commanded torque										■					
robot joint external torque									■	■					
moving camera						■		■				■	■		■
# subjects	4	5	4	4	4	20	16	4	-	-	n/a	8	4	-	2

whereas [44] specifically uses “answer phone” and “dial on a phone”, and [25] specifically uses “talk on the phone”.

IV. Datasets of Grasping

In this section, we introduce HandCorpus (`http://www.handcorpus.org`) [53] and then review 9 datasets about human hand kinematics recorded in grasping and manipulation tasks that are collected by the HandCorpus community and are stored in the HandCorpus repository. Unlike datasets [46], [47], and [50], which focus on video data, in this section we report on datasets about kinematic recording of human hand pose, in terms of sensor readings (marker positions and raw sensor data from a glove-based hand pose reconstruction system) and joint angles of human hand during reach to grasp, manipulation or grasps of real or imaginary objects.

A. The HandCorpus Initiative

HandCorpus is an open access (no login or membership required) initiative / repository for sharing datasets, tools and experimental results about human and robotic hands. The repository provides an accurate and coherent record for citing data sets, giving due credit to authors. Data sets are hierarchically indexed and can be easily retrieved using keywords and advanced search operations.

The motivation for HandCorpus is to provide the multi-disciplinary ”hand community” with a tool for benchmarking, results re-using and to foster collaborations between scientistics in the fields of neuroscience and robotics. Of course, we study hands to understand the language of the human embodiment but also to try to reproduce this incredible language under a technological point of view. Under these considerations, the importance to have a common scientific framework is crucial.

The HandCorpus was originally created in 2011, within the European Project “The Hand Embodied (THE)”, with the objective of making data collections and analyses about human hand publicly available. Since then, the HandCorpus website and structure have been ameliorated; today (July 2016), the HandCorpus repository contains 9 datasets about human hand kinematics recorded in grasping and manipulation tasks, with real or imaginary objects and 2 descriptions of kinematic models of human hands.

Regarding the latter point, it is possible to find : (i) the description of a kinematic human hand model [61] devised from Magnetic Resonance Imaging of a female hand, which provides axis locations in the form of transformation matrices, endowed with a visualization tool for the hand skeleton developed in Opensim (`http://opensimulator.org//`), (a freely available, user extensible software system to develop

TABLE VI
SHARED ANNOTATED ADLS.

Activities	[38]	[44]	[28]	[25]	[26]	[49]	[41]
use phone		answer phone, dial on a phone		talk on the phone		■	■
write on whiteboard		■		■			
drink		■	sip	■	■	■	■
eat		■			■		■
chop/cut		chop	cut	chop			
reach	■		■		■		
release	release grasp		■				
comb hair						■	■
brush teeth				■		■	■
use computer				■		■	
move					■	dishes	
stir			■	■			
pour					■		■
open	door, drawer		■		■		
close	door, drawer		■		■		

We only consider low-level annotations for [28].

models of musculoskeletal structures and create dynamic simulations of movement); (ii) the 3D coordinates of a static pose of a human hand using a motion capture system and the kinematic model described in [42]. The coordinates are expressed in C3D format (`https://www.c3d.org/`) that represents a 3D biomechanics data standard.

In addition, HandCorpus contains 5 datasets about robotic hands. There are 3 entries on the description of device architecture and specs, as well as links to schematics and files to *in house* build the robotic devices. The OpenBionics under-actuated, compliant and modular robot hands [62] and the OpenBionics light-weight, affordable, anthropomorphic prosthetic hand [63] have been developed by the OpenBionics Initiative and all the files required for their replication are available at the OpenBionics website (`http://www.openbionics.org/`). The OpenBionics is an open-source initiative for the development of affordable, light-weight, modular robotic and bionic devices founded in 2013. The Pisa/IIT SH [19] is an affordable anthropomorphic robot hand, which embeds within its design the concept of kinematic synergies [13] to move according to the first human principal grasping pattern in free motion, and adaptability, which enables the hand to deform with the external environment in order to grasp a large variety of objects. The hand has 19 joints, but only uses 1 actuator, and it is very soft and safe, yet powerful and extremely robust. The schematics of the SH can be retrieved from a direct link to Natural Machine Motion Initiative (NMMI) (`http://www.naturalmachinemotioninitiative.com/`). The NMMI is a modular open platform aiming to provide the scientific community with tools for fast and easy prototyping of Soft Robots, such as variable stiffness actuators, soft grippers, a pool of application specific add-ons, an open mechanical standardized interconnection system and common open electronics and software infrastructure to enable system integration. The Pisa/IIT SoftHand has also served as the starting point for the development of affordable and easy-to-use prostheses, whose realization is actually pursued within the EU-H2020 funded grant SoftPro – the latter is also one of the sponsors of HandCorpus together with other 6 European grants (see Acknowledgments) and the 22 international research groups across Europe, Asia and United States of America forming the HandCorpus community.

HandCorpus also contains kinematic recordings from two robotic hands: i) the fingertip positions of RBO Hand 2 [64] while enacting the Feix grasps [65] with the list of 3D coordinates of each of the five fingers, and the joint angles of a robot hand (schunk DLR HIT 4 fingers), while being tele-operated by a human hand wearing a Cyberglove [66]. The intention was to capture the workspace of the robotic hand, while avoiding possible collisions so as for the workspace to be modelled with the concept of principal motion directions, thus providing a reduced space for motion planning.

Finally, HandCorpus contains tools for the analysis, visualization and study of human and robot hands, including psychophysical investigation, tactile sensing and biomechanical modeling. As previously mentioned, HandCorpus is also a hub to other open-access initiatives about robotic and human hands. A blog, a newsletter, a publication repository and HandCorpus profiles in all major social networks are also provided. For further information, please visit the HandCorpus website, which is cross-platform, cross-browser and easily accessible through mobile devices, such as internet enabled smartphones and tablets.

In Table VII, we have reported an overview of the datasets about human hand kinematics included in the HandCorpus, with a description of the different labels used to characterize data. Focusing on human hand kinematics, the following 9 datasets refer to grasping and activities of daily living, such as haptic exploration.

B. DLR Dataset

The DLR dataset (May 2012) [13] contains the kinematics of the human hand - joint angles (captured with a passive-marker based motion capture system - Vicon), while executing the grasps reported in [13]. The results are different from [13], since the objects grasped are real and the contact forces between the fingers and the object surface induce certain deformations to the hand postures. Data of seven subjects were included and twenty three different objects were grasped.

C. HUST Dataset

The HUST Dataset (March 2016) [31] reports the joint angles of the human hand while executing the grasping tasks of the Feix taxonomy [65]. During the experiments, the subjects (5) were seated and they had their right arm fixed on the table surface in a comfortable posture. The subjects were instructed to perform thirty three types of tasks of the Feix Taxonomy [65], using a large number of objects. The human hand motion was captured with a dataglove (Cyberglove system).

D. NTUA Dataset

The NTUA Dataset (May 2010) [33] investigates the role of hand synergies during reach to grasp. A subject was seated on a chair, while his trunk was restrained to the chair and his hand was placed on the table with the palm facing downwards. Objects of varying shape and size were placed at a higher point than the starting hand position. The user was instructed to move his arm in order to reach and grasp the object. For each trial the starting hand position and the object position were kept the same. The human hand kinematics was described in terms of joint angles and captured with a Cyberglove system.

E. TU Berlin Dataset

The TU Berlin Dataset 1 (June 2015) [36] contains sensor raw data of five participants enacting the grasps of the Feix taxonomy (thirty three grasps) [65]. During the resting periods, the subjects were asked to place the hand on a table surface. The human motion was captured with a Cyberglove.

F. UNIPI-ASU Dataset

The UNIPI-ASU Dataset (May 2011) [13] reports the joint angles of the human hand, while grasping fifty seven imaginary objects according to the procedure proposed by Santello et al [13]. Human hand motion was once again captured with a Cyberglove.

G. UNIPI Datasets

The UNIPI Dataset (October 2011) [39] contains the joints angles of fifty seven grasps of imaginary objects [13], captured with an optical motion capture system (Phase Space). A single subject (male, 26) participated in the experiments.

The UNIPI Dataset 2 (September 2013) [42] contains the joint angles of the human hand of a female right handed subject described according to the model reported in [42], while grasping several imaginary objects, and recorded through a marker-based motion capture system (Phase Space). The subject was comfortably seated with the flat hand on the leg and was asked to move the hand so as to grasp an imaginary object for tool use, and then return to the rest position.

The UNIPI Dataset 3 (June 2014) [42] contains markers coordinates of a human hand of a single subject acquired through the Phase Space system, while executing the Kapandjii movement [67]. The kinematic model under investigation was described in [42].

The UNIPI Dataset 4 (September 2015) [42] contains data of a single subject that was blindfolded and was asked to haptically identify some common objects. Before each trial, the subject placed the dominant hand on a table and one of the objects was placed in random order about 30 cm in front of the hand. On a go signal, the subject reached out, explored and identified the object. In addition, the subject was also asked to explore the surface curvature, edges and textures of each object in order to prolong the exploration time. The human hand kinematics was captured with a Phase Space system and data represents joint angles.

H. Summary

From Table VII we can observe that human hand kinematics can be acquired through different acquisition systems (active marker-based motion capture system, passive-marker motion capture system (Vicon) and data glove), and using different descriptors (raw sensor data, joint angles, marker 3D coordinates). The most common acquisition modalities are (i) active marker motion capture system and (ii) data glove. The main reason for this relies on the high accuracy (the amount of static marker jitter is inferior than 0.5 mm, usually 0.1 mm) and the ease in handling marker IDs for (i), while wearability and the fact that there is no need for implementing filtering techniques to reconstruct joint angles from marker measurement represents the main motivations for using (ii). Regarding data descriptors, joint angles represent the most common type of data. This is intuitive, since this information is crucial for grasp planning [54], to drive the design and control of robotic hands [12], [19] and to improve the performance of hand-pose reconstruction systems, see e.g [68], [69].

Data format also varies (mainly .txt, .dat and .mat, but also .C3D and .csv), although .mat and .txt are the most used ones. Main motivations for this are simplicity (.txt) and the fact that joint angles are usually obtained after a post-processing phase, which is commonly performed in Matlab or Mathematica (.mat).

Finally, regarding the type of actions, reach to grasp and grasp of real objects are the most represented within HandCorpus, although grasping of imaginary objects, haptic exploration, free hand motion can be also found.

Conclusion that can be drawn is that a standardized procedure for data acquisition and data format still lacks, and it would be needed to facilitate data re-usage (e.g. the usage

TABLE VII
HANDCORPUS STATISTICS

Attributes	[13]	[31]	[33]	[36]	[13]	[39]	[42]	[42]	[42]
human postures	■	■	■	■	■	■	■	■	■
joint angles	■	■	■		■	■	■		■
marker coordinates								■	
joint sensor raw data				■					
static grasps	■		■	■	■	■			
reach & grasp	■	■	■	■			■		
free space							■	■	
haptic exploration									■
active marker motion capture system						■	■	■	■
passive marker motion capture system	■								
cyberglove		■	■	■	■				
objects type	R	R	R	R	I	I	I	N	R
# of DOFs	20	16	20	23	15	15	24	24	26
# of subjects	7	30	1	5	1	1	1	1	1
year (20–)	12	16	10	15	11	11	13	14	15

In "objects type", "R" stands for "real", "I" stands for "imaginary" and "N" for "no object".

of .C3D biomechanics standard). However, there is a clear trend in favor of the employment of motion capture and glove-based systems, joint angles as type of data, .mat and .txt for data format. Under this regard, it would be important to increase the information available on the kinematic model in use, with schemes and visual representations provided together with datasets, thus enabling a correct and simple re-usage and interpretation of data. This is already partly done within HandCorpus thanks to the usage of accompanying *Read me* file for the datasets, but it could be further improved through the adoption of common and unique data descriptions.

V. DISCUSSION

Table IX lays out different modalities included in the datasets of all three categories. We can clearly see that the different focuses require different modalities. Twenty out of twenty-two datasets in cooking/ADL tasks have RGB videos, but none of the grasping datasets has any video except for the UNIPI dataset 3, where a video of the rendered skeleton of the hand while performing grasping actions is also provided and is freely available in the dedicated YouTube channel of the HandCorpus initiative `https://youtu.be/wTZdAFGjHpI`. Such a visualization is important as it practically demonstrates the actions under investigation and increases the data comprehension. All nine datasets in the grasping category contain hand tracking data in contrast with the cooking and ADL categories that contain only one or two datasets with hand tracking data combined. Seven of nine datasets in grasping have also joint/skeleton angles, comparing with two out of eight datasets in ADL tasks and zero out twelve datasets in cooking tasks.

Research in object manipulation might find 3D object poses very useful. Explicit or readily usable recordings of object poses are available in [48]. Poses of the robot end effector are provided in [43] and [37]. [30] provides *estimated* object pose trajectories. Object poses may be computed using acceleration and rotational velocity, and object motions that are simpler than poses can be obtained if a sensor actively takes samples and is attached to an object. Datasets with such setup include

1) [27] and [29] where objects were equipped with accelerometers,
2) [28] where objects were equipped with accelerometers and rotational velocity sensors. Furniture and appliances were equipped with reed switches and accelerometers,
3) [38] where doors were equipped with reed switches.

The shared activities demonstrate a consensus among different authors on what activities should be performed and annotated. For example, certain grasping taxonomies are often adopted and such directions can be helpful for one who tries to create a new dataset. However, not being a commonly shared activity does not necessarily mean an activity is not important. Therefore, we also provide the complete list of annotated activities at `http://rpal.cse.usf.edu/motiondatasetreview/index.htm`, for cooking and ADL, respectively. The shared activities can also help with using more than one dataset. If one wants to study a certain shared activity, one could use several datasets that include this activity in order to access more modalities and higher variability. Objects that are involved in an activity may also be helpful for activity analysis. For all datasets except [34], objects are identifiable in the annotated activities through

1) being separately annotated: [23], [24], [40], [26], [49], [46], [45],
2) being part of the annotation phrases: [27], [32], [21], [21]+, [22], [29], [30], [35], [38], [44], [25], [41], [45],
3) being equipped with sensors
   a) accelerometers: [27], [29], [28],
   b) rotational velocity sensors: [28],
   c) reed switches: [38], [28],

d) RFID: [32], [38].

Temporal segmentation of annotated activities is also important for activity analysis. For [46] [50], [47], temporal segmentation does not apply because they focus on grasp instances. All other datasets include temporal segmentation, in the following forms

1) video subtitle: [27], [30],
2) explicit video time: [21]+, [49],
3) frame number: [32], [21], [22], [23], [24], [34], [40], [35], [38], [26], [45],
4) timestamp: [29], [28],
5) implicit: [44], [25], [41].

As previously mentioned in the Introduction for the case of the whole-body motion datasets, we are aware of the existence of other related datasets, however, to keep this work focused we do not include them. Examples of the excluded datasets are [1]

1) [71], and [72], [73], [74], which are datasets that do not include object manipulation motions, or if they do, the object manipulation motions are sparse,
2) [75], [76], and [3], which are dataset of objects that are typically involved in manipulation, rather than datasets of motion.

Most datasets are intended for action recognition. However, researchers who work on learning from demonstration (LfD) [55] intend to reproduce human actions rather than recognizing them. Thus, we suggest in addition to choosing from the modalities we have reviewed, a more ideal dataset for LfD should also aim to provide readily usable data that are more closely related to dynamic and kinematic motion execution. Examples of suggested modalities include trajectories of object poses, joint poses of human upper body, hand posture, torque, force between hand and object, etc.

Finally, an important specification for creating useful datasets that can be used in robotics applications, is to facilitate benchmarking. One interesting example is provided in [3], where the objects used for manipulation were chosen to cover different aspects of the manipulation problem and object characteristics, and RGB-D object scans, physical properties and geometric models are also provided together with protocol examples and physical object delivery.

VI. Conclusions

We reviewed twenty-eight datasets on object manipulation and nine datasets on grasping. We reported the characteristics and modalities of each dataset individually, we gave our view on the relation between each dataset and object manipulation, and we compared and summarized all of them together.

The datasets were created to serve their own purposes and many of them are unique. Therefore different modalities were used. The modalities range from popular video recording to rarely used air temperature and light. Many datasets were collected with numerous subjects, while some were collected with only one subject. Several datasets provide motion annotations. Twenty-three different cooking-related motions and 15 different ADL motions are annotated in the examined datasets. The survey provides a "map" for researchers in choosing the right existing dataset(s) for their own research purposes. If the right datasets are not found, the researchers may decide on creating new datasets that will supplement the exiting datasets. For example, we have not come across a dataset that includes interactive force or torque.

Observing the diversity of the datasets, we understand that trying to get a unique standard for the different types of datasets is clearly a daunting and challenging task. However, moving towards a common standardization that defines common data formats for common working areas as well as acquisition protocols would enable efficient data re-usage and sharing, fostering collaborations, and creating large datasets that allow big-data-driven approaches such as deep learning. It has been discussed recently in many conferences and workshops of the robotics community as one of several important initiatives. HandCorpus as one example is exploring a central depository approach, by collecting and processing existing datasets to provide consistent data formats and guarantee data quality.

This survey does not include datasets that, although are introduced in publications, are not openly available. Many of them were presented in the Workshop on Grasping and Manipulation Datasets that was organized under the International Conference on Robotics and Automation (ICRA) in May 2016. The workshop's report [77] provides a survey of those works and datasets.

VII. Acknowledgement

This work was partly supported by the National Science Foundation under Grants No. 1421418 and No. 1560761, by the European Research Council under the Advanced Grant SoftHands (No. ERC-291166), by the EU H2020 projects SoftPro (No. 688857) and SOMA (No. 645599), and by the EU FP7 project WEARHAP (No. 601165), PACMAN (No. 600918), WALKMAN (No. 611832) and THE (No. 248587).

References

[1] C. Mandery, Ö. Terlemez, M. Do, N. Vahrenkamp, and T. Asfour, "The kit whole-body human motion database," in *Advanced Robotics (ICAR), 2015 International Conference on*. IEEE, 2015, pp. 329–336.

[2] Y. Sun, Y. Lin, and Y. Huang, "Robotic grasping for instrument manipulations," in *International Conference on Ubiquitous Robots and Ambient Intelligence 2016*, 2016, pp. 1–3.

[3] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar, "The ycb object and model set: Towards common benchmarks for manipulation research," in *proceedings of the 2015 IEEE International Conference on Advanced Robotics (ICAR)*, 2015.

[1] HandCorpus represents an interesting example of datasets whose role can bridge the gap between neuroscience and robotics. However, in literature, it is possible to find also datasets specifically built to address purely neuroscientific questions, which are currently out of the scope of this review. It is the case of the WAY-EEG-GAL [70], which was designed to allow critical tests of techniques to decode sensation, intention, and action from scalp EEG recordings in humans who perform a grasp-and-lift task. Twelve participants performed lifting series (a total of 3,936 trials) in which the object's weight, surface friction, or both, were changed unpredictably between trials. EEG, EMG, the 3D position of both the hand and object (through the Pholemus passive-marker magnetic motion capture system), as well as force/torque at both contact plates were recorded.

[4] M. Kopicki, R. Detry, F. Schmidt, C. Borst, R. Stolkin, and J. L. Wyatt, "Learning dexterous grasps that generalise to novel objects by combining hand and contact models," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2014, pp. 5358–5365.

[5] Y. Bekiroglu, D. Kragic, and V. Kyrki, "Learning grasp stability based on tactile data and hmms," in *19th International Symposium in Robot and Human Interactive Communication*. IEEE, 2010, pp. 132–137.

[6] J. Bohg, A. Morales, T. Asfour, and D. Kragic, "Data-driven grasp synthesis—a survey," *IEEE Transactions on Robotics*, vol. 30, no. 2, pp. 289–309, 2014.

[7] J. Mahler, F. T. Pokorny, B. Hou, M. Roderick, M. Laskey, M. Aubry, K. Kohlhoff, T. Kröger, J. Kuffner, and K. Goldberg, "Dex-net 1.0: A cloud-based network of 3d objects for robust grasp planning using a multi-armed bandit model with correlated rewards."

[8] M. Bianchi and A. Moscatelli, "Human and robot hands."

[9] D. Prattichizzo and J. C. Trinkle, "Grasping," in *Springer handbook of robotics*. Springer, 2008, pp. 671–700.

[10] Y. Lin and Y. Sun, "Task-based grasp quality measures for grasp synthesis," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 485–490.

[11] ——, "Grasp planning to maximize task coverage," *The International Journal of Robotics Research*, vol. 34, no. 9, pp. 1195–1210, 2015.

[12] M. Santello, M. Bianchi, M. Gabiccini, E. Ricciardi, G. Salvietti, D. Prattichizzo, M. Ernst, A. Moscatelli, H. Jörntell, A. M. Kappers *et al.*, "Hand synergies: integration of robotics and neuroscience for understanding the control of biological and artificial hands," *Physics of life reviews*, 2016.

[13] M. Santello, M. Flanders, and J. F. Soechting, "Postural hand synergies for tool use," *The Journal of Neuroscience*, vol. 18, no. 23, pp. 10 105–10 115, 1998. [Online]. Available: http://www.jneurosci.org/content/18/23/10105.abstract

[14] J. F. Soechting and M. Flanders, "Flexibility and repeatability of finger movements during typing: analysis of multiple degrees of freedom," *Journal of computational neuroscience*, vol. 4, no. 1, pp. 29–46, 1997.

[15] F. Ficuciello, G. Palli, C. Melchiorri, and B. Siciliano, "Experimental evaluation of postural synergies during reach to grasp with the ub hand iv," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2011, pp. 1775–1780.

[16] M. Ciocarlie, C. Goldfeder, and P. Allen, "Dimensionality reduction for hand-independent dexterous robotic grasping," in *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2007, pp. 3270–3275.

[17] M. T. Ciocarlie and P. K. Allen, "Hand posture subspaces for dexterous robotic grasping," *The International Journal of Robotics Research*, vol. 28, no. 7, pp. 851–867, 2009.

[18] C. Y. Brown and H. H. Asada, "Inter-finger coordination and postural synergies in robot hands via mechanical implementation of principal components analysis," in *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2007, pp. 2877–2882.

[19] M. G. Catalano, G. Grioli, E. Farnioli, A. Serio, C. Piazza, and A. Bicchi, "Adaptive synergies for the design and control of the pisa/iit softhand," *The International Journal of Robotics Research*, vol. 33, no. 5, pp. 768–782, 2014.

[20] S. Ciotti, E. Battaglia, N. Carbonaro, A. Bicchi, A. Tognetti, and M. Bianchi, "A synergy-based optimally designed sensing glove for functional grasp recognition," *Sensors*, vol. 16, no. 6, p. 811, 2016. [Online]. Available: http://www.mdpi.com/1424-8220/16/6/811

[21] A. Fathi, Y. Li, and J. M. Rehg, "Learning to recognize daily actions using gaze," in *Proceedings of the 12th European Conference on Computer Vision - Volume Part I*, ser. ECCV'12, 2012, pp. 314–327.

[22] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele, "A database for fine grained activity detection of cooking activities," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012.

[23] M. Rohrbach, M. Regneri, M. Andriluka, S. Amin, M. Pinkal, and B. Schiele, "Script data for attribute-based recognition of composite activities," in *European Conference on Computer Vision (ECCV)*, October 2012.

[24] 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," 2015.

[25] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Human activity detection from rgbd images," in *In AAAI workshop on Pattern, Activity and Intent Recognition (PAIR)*, 2011.

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

[27] C. Pham and P. Olivier, *Slice&Dice: Recognizing food preparation activities using embedded accelerometers*. Springer, 2009, pp. 34–43.

[28] D. Roggen, A. Calatroni, M. Rossi, T. Holleczek, K. Forster, G. Troster, P. Lukowicz, D. Bannach, G. Pirkl, A. Ferscha, J. Doppler, C. Holzmann, M. Kurz, G. Holl, R. Chavarriaga, H. Sagha, H. Bayati, M. Creatura, and J. del R Millan, "Collecting complex activity datasets in highly rich networked sensor environments," in *Networked Sensing Systems (INSS), 2010 Seventh International Conference on*, June 2010, pp. 233–240.

[29] S. Stein and S. J. McKenna, "Combining embedded accelerometers with computer vision for recognizing food preparation activities," in *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2013, pp. 729–738.

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

[31] M.-J. Liu, C.-H. Xiong, L. Xiong, and X.-L. Huang, "Biomechanical characteristics of hand coordination in grasping activities of daily living," *PLoS ONE*, vol. 11, no. 1, pp. 1–16, 01 2016. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.pone.0146193

[32] F. de la Torre, J. Hodgins, A. Bargteil, A. Collado, X. Martin, J. Macey, and P. Beltran, "Guide to the carnegie mellon university multimodal activity (cmu-mmac) database," Robotics Institute, Carnegie Mellon University, Tech. Rep. CMU-RI-TR-08-22, July 2009.

[33] M. V. Liarokapis, P. K. Artemiadis, and K. J. Kyriakopoulos, "Telemanipulation with the dlr/hit ii robot hand using a dataglove and a low cost force feedback device," in *Control Automation (MED), 2013 21st Mediterranean Conference on*, June 2013, pp. 431–436.

[34] A. Shimada, K. Kondo, D. Deguchi, G. Morin, and H. Stern, "Kitchen scene context based gesture recognition: A contest in ICPR2012," *Advances in Depth Image Analysis and Applications, Lecture Notes in Computer Science*, vol. 7854, pp. 168–185, 2013.

[35] H. Kuehne, A. Arslan, and T. Serre, "The language of actions: Recovering the syntax and semantics of goal-directed human activities," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2014.

[36] R. Deimel and O. Brock, "A novel type of compliant and underactuated robotic hand for dexterous grasping," *The International Journal of Robotics Research*, 2015. [Online]. Available: http://ijr.sagepub.com/content/early/2015/08/13/0278364915592961.abstract

[37] S. Levine, P. Pastor, A. Krizhevsky, and D. Quillen, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," *CoRR*, vol. abs/1603.02199, 2016. [Online]. Available: http://arxiv.org/abs/1603.02199

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

[39] M. Bianchi, P. Salaris, and A. Bicchi, "Synergy-based hand pose sensing: Reconstruction enhancement," *The International Journal of Robotics Research*, 2013. [Online]. Available: http://ijr.sagepub.com/content/early/2013/02/19/0278364912474078.abstract

[40] P. Das, C. Xu, R. Doell, and J. Corso, "A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 2013.

[41] B. Bruno, F. Mastrogiovanni, and A. Sgorbissa, "A public domain dataset for adl recognition using wrist-placed accelerometers," in *Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on*, 2014, pp. 738–743.

[42] M. Gabiccini, G. Stillfried, H. Marino, and M. Bianchi, "A data-driven kinematic model of the human hand with soft-tissue artifact compensation mechanism for grasp synergy analysis," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Nov 2013, pp. 3738–3745.

[43] C. Finn, I. J. Goodfellow, and S. Levine, "Unsupervised learning for physical interaction through video prediction," *CoRR*, vol. abs/1605.07157, 2016. [Online]. Available: http://arxiv.org/abs/1605.07157

[44] R. Messing, C. Pal, and H. Kautz, "Activity recognition using the velocity histories of tracked keypoints," in *ICCV*, 2009.

[45] A. Fathi, X. Ren, and J. M. Rehg, "Learning to recognize objects in egocentric activities," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, June 2011, pp. 3281–3288.

[46] I. M. Bullock, T. Feix, and A. M. Dollar, "The Yale human grasping data set: Grasp, object, and task data in household and machine shop

environments," *International Journal of Robotics Research*, vol. 34, no. 3, pp. 251–255, 2014.

[47] G. Rogez, J. S. Supancic, and D. Ramanan, "Understanding everyday hands in action from rgb-d images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3889–3897.

[48] [Online]. Available: http://rpal.cse.usf.edu/imd/

[49] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, June 2012, pp. 2847–2854.

[50] M. Cai, K. M. Kitani, and Y. Sato, "A scalable approach for discovering the visual structures of hand grasps," May 2015.

[51] T.-H. Pham, N. Kyriazis, A. A. Argyros, and A. Kheddar, "Hand-Object Contact Force Estimation From Markerless Visual Tracking," Aug. 2016, submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01356138

[52] G. Rogez, J. S. S. III, M. Khademi, J. M. M. Montiel, and D. Ramanan, "3d hand pose detection in egocentric RGB-D images," *CoRR*, vol. abs/1412.0065, 2014. [Online]. Available: http://arxiv.org/abs/1412.0065

[53] M. Bianchi and M. V. Liarokapis, "HandCorpus, a new open-access repository for sharing experimental data and results on human and artificial hands," in *IEEE World Haptics Conference (WHC)*, April 2013.

[54] H. Marino, M. Ferrati, A. Settimi, C. Rosales, and M. Gabiccini, "On the problem of moving objects with autonomous robots: A unifying high-level planning approach," *IEEE ROBOTICS AND AUTOMATION LETTERS*, vol. 1, pp. 469–476, 01/2016 2016. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7384694

[55] A. Billard, S. Calinon, R. Dillmann, and S. Schaal, "Robot programming by demonstration," in *Springer Handbook of Robotics*. Springer Berlin Heidelberg, 2008.

[56] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," *CoRR*, vol. abs/1411.4389, 2014. [Online]. Available: http://arxiv.org/abs/1411.4389

[57] D. Paulius, R. Milton, Y. Huang, W. Buchanan, J. Sam, and Y. Sun, "Functional object-oriented network for manipulation learning," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016, pp. 1–8.

[58] I. M. Bullock, T. Feix, and A. M. Dollar, "Finding small, versatile sets of human grasps to span common objects," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, May 2013, pp. 1068–1075.

[59] J. Liu, F. Feng, Y. C. Nakamura, and N. S. Pollard, "A taxonomy of everyday grasps in action," in *2014 IEEE-RAS International Conference on Humanoid Robots*. IEEE, 2014, pp. 573–580.

[60] M. Gleicher, "Retargetting motion to new characters," in *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '98. New York, NY, USA: ACM, 1998, pp. 33–42. [Online]. Available: http://doi.acm.org/10.1145/280814.280820

[61] G. Stillfried, U. Hillenbrand, M. Settles, and P. van der Smagt, "Mri-based skeletal hand movement model," in *The Human Hand as an Inspiration for Robot Hand Development*. Springer International Publishing, 2014, pp. 49–75.

[62] A. G. Zisimatos, M. V. Liarokapis, C. I. Mavrogiannis, and K. J. Kyriakopoulos, "Open-source, affordable, modular, light-weight, underactuated robot hands," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 3207–3212.

[63] G. P. Kontoudis, M. V. Liarokapis, A. G. Zisimatos, C. I. Mavrogiannis, and K. J. Kyriakopoulos, "Open-source, anthropomorphic, underactuated robot hands with a selectively lockable differential mechanism: Towards affordable prostheses," in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, Sept 2015, pp. 5857–5862.

[64] R. Deimel and O. Brock, "A novel type of compliant and underactuated robotic hand for dexterous grasping," *The International Journal of Robotics Research*, p. 0278364915592961, 2015.

[65] T. Feix, R. Pawlik, H.-B. Schmiedmayer, J. Romero, and D. Kragic, "A comprehensive grasp taxonomy," in *Robotics, Science and Systems: Workshop on Understanding the Human Hand for Advancing Robotic Manipulation*, 2009.

[66] J. Rosell, R. Suárez, C. Rosales, and A. Pérez, "Autonomous motion planning of a hand-arm robotic system based on captured human-like hand postures," *Autonomous Robots*, vol. 31, no. 1, pp. 87–102, 2011.

[67] A. Kapandji, "[clinical test of apposition and counter-apposition of the thumb]," *Annales de chirurgie de la main : organe officiel des societes de chirurgie de la main*, vol. 5, no. 1, p. 67—73, 1986. [Online]. Available: http://dx.doi.org/10.1016/S0753-9053(86)80053-9

[68] M. Bianchi, P. Salaris, and A. Bicchi, "Synergy-based hand pose sensing: Reconstruction enhancement," *The International Journal of Robotics Research (IJRR)*, vol. 32, no. 4, pp. 396–406, 2013.

[69] ——, "Synergy-based hand pose sensing: Optimal glove design," *The International Journal of Robotics Research*, vol. 32, no. 4, pp. 407–424, 2013.

[70] M. D. Luciw, E. Jarocka, and B. B. Edin, "Multi-channel eeg recordings during 3,936 grasp and lift trials with varying weight and friction," *Scientific data*, vol. 1, p. 140047, 2014.

[71] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero, "A survey of video datasets for human action and activity recognition," *Computer Vision and Image Understanding*, vol. 117, no. 6, pp. 633 – 659, 2013.

[72] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human action classes from videos in the wild," CRCV-TR-12-01, Tech. Rep., 2012.

[73] T. Huynh, M. Fritz, and B. Schiele, "Discovery of activity patterns using topic models," in *Proceedings of the 10th International Conference on Ubiquitous Computing*, ser. UbiComp '08, 2008, pp. 10–19.

[74] E. Ohn-Bar and M. Trivedi, "The power is in your hands: 3d analysis of hand gestures in naturalistic video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 912–917.

[75] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view rgb-d object dataset," in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, May 2011, pp. 1817–1824.

[76] A. Singh, J. Sha, K. Narayan, T. Achim, and P. Abbeel, "Bigbird: A large-scale 3d database of object instances," in *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, May 2014, pp. 509–516.

[77] M. Bianchi, J. Bohg, and Y. Sun, "Latest datasets and technologies presented in the workshop on grasping and manipulation datasets," *arXiv*, 2016.

TABLE VIII
Shared annotated cooking activities

Activity	[27]	[32]	[21]	[21]+	[24]	[29]	[34]	[40]	[30]	[35]	[45]
chop/cut	chop, slice, dice			cut	chop, cut, cut apart, cut dice, cut off ends, cut off inside, cut stripes, slice	cut	cut			cut	
peel/shave	peel, shave			peel	peel	peel	peel			peel	
stir/mix	stir	stir		mix	mix, stir	mix	mix	stir		stir	stir
pour		■	■	■	■			■	milk, cereal	■	■
put/place		put		put	put in, put on	place		put down		put	put
take		■	■	■	take lid, take out					■	■
spread/smear	spread		spread	spread	spread					smear	spread
eat/taste	eat				taste						
scoop/spoon	scoop		scoop							spoon	scoop
season/spice					spice		season	season			
turn/flip				flip	turn over		turn	flip			
open/close food (container)		open	■	■	■				■		■
open/close drawer		open			■						
open/close dishwasher/oven				oven	■						
open/close cupboard /fridge /microwave		■		fridge	■						
crack/break		egg		■	open egg		■			■	
beat/whip		beat egg			whip						
add					■	■				teabag, salt and pepper, topping	
squeeze				■	■					■	
turn on/off				■	■						
wash				■	■						
dry				■	■						
fill				■	■						

Since [24] supercedes [22] and [23], we only include [24] in the table.

TABLE IX
MODALITIES

	COOKING												
Modalities	[27]	[32]	[45]	[21]	[21]+[22]		[23]	[24]	[29]	[34]	[40]	[30]	[35]
RGB video (image sequences)	■	■	■	■	■	■	■	■	■	■	■	■	■
depth image sequence									■	■		■	
audio		■			■							■	
RFID		■											
motion capture		■											
glove-based system		■											
Bayer pattern video													
3D acceleration on human		■											
3D rotational velocity on human		■											
3D orientation on human													
3D location on human													
2D acceleration on human		■											
skeleton													
joint angles													
body heat		■											
3D magnetic field on human		■											
skin temperature		■											
skin conductance		■											
gaze				■	■								
3D acceleration on object	■								■				
3D rot. vel. & acc. on object													
2D rotational velocity on object													
6D object pose track												■	
object models												■	
3D acceleration on furniture													
reed switch on furniture													
force on fingertip													
in-tool force & torque													
gripper pose													
commanded gripper pose													
robot joint position													
robot joint velocity													
robot joint commanded torque													
robot joint external torque													
air temperature		■											
light		■											
moving camera		🟦	🟦	🟦	🟦						🟦		
activity type	C	C	C	C	C	C	C	C	C	C	C	C	C
# subjects	20	>394		14	10	12	30	30	25	7	n/a	8	52
year (20–)	09	09	11	12	12	12	12	15	13	13	13	14	14

	ADL														
Modalities	[38]	[44]	[28]	[25]	[26]	[49]	[41]	[46]	[43]	[37]	[48]	[47]	[50]	[51]	[52]
RGB video (image sequences)	■	■		■	■	■		■	■	■	■	■	■		■
depth image sequence				■	■					■	■	■			■
audio															
RFID	■														
motion capture	■										■				
glove-based system	■										■				
Bayer pattern video	■														
3D acceleration on human			■				■								
3D rotational velocity on human			■												
3D orientation on human			■												
3D location on human			■												
2D acceleration on human															
skeleton				■	■										
joint angles				■	■										
body heat															
3D magnetic field on human			■												
skin temperature															
skin conductance															
gaze															
3D acceleration on object			■											■	
3D rot. vel. & acc. on object														■	
2D rotational velocity on object			■												
6D object pose track											■				
object models															
3D acceleration on furniture			■												
reed switch on furniture	■		■												
force on fingertip														■	
in-tool force & torque											■				
gripper pose									■	■					
commanded gripper pose									■	■					
robot joint position									■	■					
robot joint velocity										■					
robot joint commanded torque										■					
robot joint external torque									■	■					
air temperature															
light															
moving camera						🟦		🟦				🟦	🟦		🟦
activity type	A	A	A	A	A	A	A	A	P	G	A	G	G	M	M
# subjects	4	5	4	4	4	20	16	4	-	-	n/a	8	4	-	2
year (20–)	09	09	10	11	13	12	14	14	16	16	16	15	15	16	14

	HAND CORPUS								
Modalities	[13]	[31]	[33]	[36]	[13]	[39]	[42]	[42]	[42]
RGB video (image sequences)									
depth image sequence									
audio									
RFID									
motion capture				■	■	■	■	■	■
glove-based system	■	■	■						
Bayer pattern video									
3D acceleration on human									
3D rotational velocity on human									
3D orientation on human								■	
3D location on human								■	
2D acceleration on human									
skeleton									
joint angles	■	■	■	■	■	■	■		■
body heat									
3D magnetic field on human									
skin temperature									
skin conductance									
gaze									
3D acceleration on object									
3D rot. vel. & acc. on object									
2D rotational velocity on object									
6D object pose track									
object models									
3D acceleration on furniture									
reed switch on furniture									
force on fingertip									
in-tool force & torque									
gripper pose									
commanded gripper pose									
robot joint position									
robot joint velocity									
robot joint commanded torque									
robot joint external torque									
air temperature									
light									
moving camera									
activity type	G	G	G	G	G	G	G	M	H
# subjects	7	30	1	5	1	1	1	1	1
year (20–)	12	16	10	15	11	11	13	14	15

In row “activity type”, “C” stands for “cooking”, “A” stands for “ADL”, “G” stands for “grasping”, “P” stands for “pushing”, “M” stands for “manipulation” and “H” stands for “haptic exploration”.