

Datasets on object manipulation and interaction: a survey

Yongqiang Huang and Yu Sun

Abstract—A dataset is crucial for model learning and evaluation. Choosing the right dataset to use or making a new dataset requires the knowledge of those that are available. In this work, we provide that knowledge, by reviewing twenty datasets that were published in the recent six years and that are directly related to object manipulation. We report on modalities, activities, and annotations for each individual dataset and give our view on its use for object manipulation. We also compare the datasets and summarize them. We conclude with our suggestion on future datasets.

I. INTRODUCTION

Datasets are valuable in various scientific fields because they are crucial for testing an algorithm. The demands for datasets follow the advancement of a field or the evolution of a problem, and new datasets never stopped being created. A good dataset may not only verify or deny the correctness and effectiveness of an algorithm, but may also help expose the flaws or exemplify the strength of the algorithm. To choose the good dataset, one first needs to know what datasets are available, what they include, and how they differ. Then one can decide on whether any datasets would be useful and which one or several 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 make new ones. To help one with choosing the right dataset(s) or deciding on making new datasets, we contribute a review of datasets that we consider would be useful for research on object manipulation. All the datasets were published in the recent six years, i.e., since 2009.

As the name implies, an object manipulation motion involves an object. It intends to accomplish a certain task by manipulating, or changing the position and orientation of the object. 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. We report on datasets that *focus* on object manipulation motion. Gross motions may be present in certain datasets, but do not play the dominant role.

We divide the datasets into two categories and present them separately: those that include mostly cooking activities, in section II, and those that include more general activities of daily living (ADL), in section III. In both categories, we sort the datasets in ascending chronological order. We keep datasets that belong to the same series together, and use the earliest publication year among the members for sorting. Fig.

The authors are with the Department of Computer Science and Engineering at the University of South Florida, Tampa, FL, USA. email: yongqiang@mail.usf.edu, yusun@cse.usf.edu.

1 shows the year of each dataset in the presented order, in which the series are delimited in green: ([3], [3]+), ([4], [5], [6]), and ([15], [16]).

We downloaded each dataset and verified the consistency of the contents with the publication. When we encountered confusion or uncertainty about certain contents, we did not assume or guess but asked the authors for clarification. For each dataset, we report on the modalities, the activities performed, and annotations, then we give our view on how the 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.

Because of the limitation in space, this review cannot be exhaustive in width or depth. To learn about datasets on more general human actions, one is directed to [20]. For those who wants to learn more about certain datasets covered in this work, we provide the link to each dataset in Table II.

dataset	[1]	[2]	[3]	[3]+	[4]	[5]	[6]	[7]	[8]	[9]
year (20--)	09	09	12	12	12	12	15	13	13	13
dataset	[10]	[11]	[12]	[13]	[14]	[15]	[16]	[17]	[18]	[19]
year (20--)	14	14	09	09	10	11	13	12	14	14

Fig. 1. The chronological order in which the datasets are presented

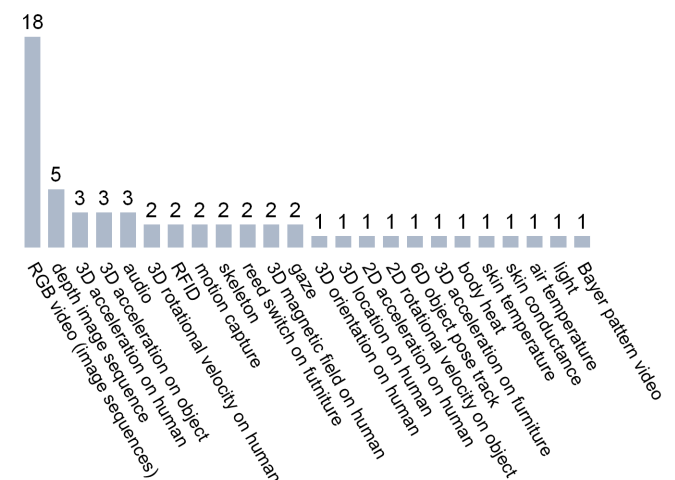


Fig. 3. Count of datasets for each modality

Modality	[1]	[2]	[3]	[3]+	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[13]	[14]	[15]	[16]	[17]	[18]	[19]
RGB video (image sequences)	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
depth image sequence								■	■		■					■	■			■
3D acceleration on human		■													■					■
3D rotational velocity on human		■													■					
3D orientation on human															■					
3D location on human															■					
2D acceleration on human		■																		
3D acceleration on object	■							■							■					
2D rotational velocity on object															■					
audio		■		■							■									
RFID		■											■							
motion capture		■											■							
skeleton																■	■			
6D object pose track											■									
3D acceleration on furniture															■					
reed switch on furniture														■						
3D magnetic field on human		■													■					
body heat		■																		
skin temperature		■																		
skin conductance		■																		
air temperature		■																		
light		■																		
gaze			■	■																
object models											■									
Bayer pattern video														■						

Fig. 2. Modalities

II. DATASETS OF COOKING ACTIVITY

A. *Slice&Dice*

Slice&Dice [1] 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 each 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 for, unlike images, acceleration data belong to a certain object alone, and is readily usable without running object recognition first.

B. *CMU-MMAC*

The *CMU-MMAC* dataset [2] contains multi-modal cooking activities of five recipes: brownie, eggs, pizza, salad, and 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 performed all the recipes. The dataset also specifically recorded anomalous accidental events that happened while cooking. Certain modalities are incomplete for certain recipes performed by certain subjects. Annotations exist for sixteen subjects making 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 *Slice&Dice* [1]. 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. *Gaze and Gaze+*

The *Gaze* dataset [3] contains RGB egocentric videos of fourteen subjects making 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 [3] (later referred to as [3]+) 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 performed 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.

D. *The MPII Cooking dataset, Cooking Composite dataset, and Cooking 2 dataset*

MPII sequentially created three datasets related to cooking: the *MPII Cooking dataset* [4] which focuses on fine grained activity, the *MPII Cooking Composite dataset* [5]

which focuses on composite activities composed of basic-level activities, and the MPII Cooking 2 dataset [6] which unifies and is an upgrade of both [4] and [5].

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 exists 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 [4] and [5], 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 [4], and forty-five shorter and simpler composite dishes from [5]. 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 [11], if not unmatched.

E. 50 Salad

The 50 Salad dataset [7] extends Slice&Dice [1] by using accelerometers on more utensils and by including depth videos in addition to RGB ones. Twenty-five subjects 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 [1], involves more subjects, enables 3D analysis with depth videos, and has finer annotations. In that regard, we recommend 50 Salad over Slice&Dice.

F. The Actions for Cooking Eggs dataset (ACE)

The ACE dataset [8] 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.

G. The YouCook dataset

The YouCook dataset [9] 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.

H. The dataset of actions for making cereal

This dataset [10] recorded eight subjects making 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.

I. The Brown breakfast actions dataset

The Brown breakfast dataset [11] 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, roughly 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.

III. DATASETS OF ACTIVITIES OF DAILY LIVING (ADL)

A. The TUM Kitchten dataset

The TUM Kitchen dataset [12] 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

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 [2], the dataset identifies 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. The Rochester ADL dataset

The Rochester ADL dataset [13] 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 [4]-[6] and Brown breakfast dataset [11], the Rochester ADL dataset would rely on human and object recognition to be useful for 2D fine motion analysis.

C. The OPPORTUNITY dataset

The OPPORTUNITY dataset [14] contains multi-modal data of five morning ADL runs and one Drill run for each of four subjects. Motion sensors were densely deployed on human body, on the objects, and in the environment. The modalities on human body include IMUs, 3D accelerometers, and 3D localizers. The modalities on objects include 3D accelerometers and 2D rotational velocity sensors. The annotations consists of five “tracks”: locomotion, high-level activity, 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* hand and objects. Since object manipulation analysis focuses on the interaction between hand and objects, data that include the motion of both hand and objects are desired. The dataset is comparable with 50 Salad [7], CMU-MMAC [2], and TUM Kitchen [12] 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. The Cornell CAD-60 and CAD-120 datasets

The CAD-60 [15] and the CAD-120 [16] 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 is similar to motion capture, but with much fewer defined joints and less corresponding data. Despite the “lightness” compared with motion capture, the skeleton is directly usable for 3D fine motion analysis, and therefore we consider it as the merit of the datasets.

E. The first person ADL dataset

The ADL dataset by Pirsivash [17] 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(+) [3], 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. The wrist-worn accelerometer dataset

The wrist-worn accelerometer dataset [18] 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. The Yale human grasping dataset

The Yale human grasping dataset [19] 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.

IV. DATASET SUMMARY

Fig. 2 lays out the different modalities included in each dataset, and Fig. 3 shows in descending order the count of datasets for each modality. We can see from the figures that, as the most easily managed modality, RGB video leads with eighteen datasets excluding only [14] and [18]. In fact, [14] did collect RGB videos but did not publish them. Depth video is the second most adopted modality, but with a count much lower than that of RGB video. 3D acceleration (on human or object) leads among the rest modalities, but no modalities besides videos stand out. Motion capture data are only found in [2] and [12], possibly because of the cost and effort required in the setup of the system (although [12] uses a markerless capture system and most of the effort is with

TABLE I
SHARED ANNOTATED ADLS.

Activities	[12]	[13]	[14]	[15]	[16]	[17]	[18]
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 [14].

computing). The skeleton tracks, which can be considered as a light version of the motion capture, are only available in [15] and [16].

Research in object manipulation might find 3D object poses very useful. Acceleration and rotational velocity may be used to estimate object poses, but explicit or readily usable recordings of object poses, which may require a motion capture system, are unavailable in the datasets. [10] is the only dataset that provides something close: the *estimated* 6 DOF object pose trajectories. 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) [1], [7]. Objects were equipped with accelerometers.
- 2) [14]. Objects were equipped with accelerometers and rotational velocity sensors. Furniture and appliances were equipped with reed switches and accelerometers.
- 3) [12]. Doors were equipped with reed switches.

Activity annotations can be useful for various purposes. We identified the annotated activities that are shared by multiple datasets, and list those belonging to the ADL datasets in Table I, and those belonging to the cooking datasets in Table III. In both tables, we combine similar annotations and specify each in the cells. For example, on the first row of Table I, the annotated activity is summarized as “use phone”, whereas [13] specifically uses “answer phone” and “dial on a phone”, and [15] specifically uses “talk on the phone”.

The shared activities show the consensus among different authors on what activities should be performed and annotated, which can be helpful for one who tries to make such decision when making a new dataset. However, because the amount of authors is limited, not being a shared activity does not necessarily mean the 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 together to access more modalities and higher variability.

Except for annotated activities, objects that are involved in an activity may also be helpful for activity analysis. For all datasets except [8], objects are identifiable in the annotated activities through

- 1) being separately annotated: [5], [6], [9], [16], [17], [19].
- 2) being part of the annotation phrases: [1], [2], [3], [3]+, [4], [7], [10], [11], [12], [13], [15], [18].
- 3) being equipped with sensors
 - a) accelerometers: [1], [7], [14].
 - b) rotational velocity sensors: [14].
 - c) reed switches: [12], [14].
 - d) RFID: [2], [12].

Temporal segmentation of annotated activities is also important for activity analysis. For [19], temporal segmentation does not apply because [19] focuses on grasp instances. All other datasets include temporal segmentation, in the following forms

- 1) video subtitle: [1], [10].
- 2) explicit video time: [3]+, [17].
- 3) frame number: [2], [3], [4], [5], [6], [8], [9], [11], [12], [16].
- 4) timestamp: [7], [14]
- 5) implicit: [13], [15], [18].

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) [20], and [21], [22], which are datasets that do not include object manipulation motions, or if they do, the object manipulation motions are sparse.
- 2) [23], [24], and [25], which are dataset of objects that are typically involved in manipulation, rather than datasets of motion.

Most datasets we include are intended for action recognition. However, researchers who work on learning from demonstration (LfD) [26] intend to reproduce human actions rather than recognizing them. Thus, we suggest that except for 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.

V. CONCLUSION

We reviewed twenty datasets that we considered useful for research on object manipulation. We reported on each dataset individually, gave our view on the relation between each dataset and object manipulation, and compared and summarized all of them together. We provided suggestion on future datasets, and we are putting that suggestion into practice and making a new dataset.

VI. ACKNOWLEDGEMENT

This material is based upon work supported by the National Science Foundation under Grant No. 1421418.

REFERENCES

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

[2] 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-mmact) database,” Robotics Institute, Carnegie Mellon University, Tech. Rep. CMU-RI-TR-08-22, July 2009.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

TABLE II
LINK TO DATASETS

[1]	http://openlab.ncl.ac.uk/publicweb/publicweb/AmbientKitchen/KitchenData/Slice&Dice_dataset/
[2]	http://kitchen.cs.cmu.edu/
[3](+)	http://ai.stanford.edu/~alireza/GTEA_Gaze_Website/
[4]	https://www.mpi-inf.mpg.de/departments/computer-vision-and-multimodal-computing/research/human-activity-recognition/mpii-cooking-activities-dataset/
[5]	https://www.mpi-inf.mpg.de/departments/computer-vision-and-multimodal-computing/research/human-activity-recognition/mpii-cooking-composite-activities/
[6]	https://www.mpi-inf.mpg.de/departments/computer-vision-and-multimodal-computing/research/human-activity-recognition/mpii-cooking-2-dataset/
[7]	http://cvip.computing.dundee.ac.uk/datasets/foodpreparation/50salads/
[8]	http://www.murase.m.is.nagoya-u.ac.jp/KSCGR/
[9]	http://www.cse.buffalo.edu/~jcorso/r/youcook/
[10]	http://robocoffee.org/datasets/
[11]	http://serre-lab.clps.brown.edu/resource/breakfast-actions-dataset/
[12]	https://ias.in.tum.de/software/kitchen-activity-data
[13]	http://www.cs.rochester.edu/~rmessing/uradl/
[14]	UCI repository: https://archive.ics.uci.edu/ml/datasets/OPPORTUNITY+Activity+Recognition# , Challenge: http://www.opportunity-project.eu/challengeDataset
[15][16]	http://pr.cs.cornell.edu/humanactivities/data.php
[17]	http://people.csail.mit.edu/hpirsiav/codes/ADLdataset/adl.html
[18]	https://archive.ics.uci.edu/ml/datasets/Dataset+for+ADL+Recognition+with+Wrist-worn+Accelerometer#
[19]	http://www.eng.yale.edu/grablab/humangrasping/

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

[14] D. Roggen, A. Calatroni, M. Rossi, T. Holleczeck, K. Forster, G. Troster, P. Lukowicz, D. Bannach, G. Pirkil, 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.

[15] 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.

- [16] 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.
- [17] 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.
- [18] 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.
- [19] 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.
- [20] J. M. Chaquet, E. J. Carmona, and A. Fernndez-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.
- [21] 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.
- [22] 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.
- [23] 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.
- [24] 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.
- [25] 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.
- [26] A. Billard, S. Calinon, R. Dillmann, and S. Schaal, "Robot programming by demonstration," in *Springer Handbook of Robotics*. Springer Berlin Heidelberg, 2008.

TABLE III
SHARED ANNOTATED COOKING ACTIVITIES

Activity	[1]	[2]	[3]	[3]+	[6]	[7]	[8]	[9]	[10]	[11]
chop/cut	chop, dice	slice,		cut	chop, 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
pour		•	•	•	•			•	milk, cereal	•
put/place		put		put	put in, put on	place		put down		put
take		•	•	•	take lid, take out					•
spread/smear	spread		spread	spread	spread					smear
eat/taste	eat				taste					
scoop/spoon	scoop		scoop							spoon
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 [6] supercedes [4] and [5], we only include [6] in the table.