A Dataset of Daily Interactive Manipulation

Yongqiang Huang and Yu Sun

Abstract
Robots that succeed in factories stumble to complete the simplest daily task humans take for granted, for the change of environment makes the task exceedingly difficult. Aiming to teach robots perform daily interactive manipulation in a changing environment using human demonstrations, we collected our own data of interactive manipulation. The dataset focuses on position, orientation, force, and torque of objects manipulated in daily tasks. The dataset includes 1,593 trials of 32 types of daily motions and 1,596 trials of pouring alone, as well as helper code. We present our dataset to facilitate the research on task-oriented interactive manipulation.

Keywords
dataset, interactive manipulation, force, motion tracking

1 Introduction
Robots excel in manufacturing which requires repetitive motion with little fluctuation between trials. In contrast, humans rarely complete any daily task by repeating exactly what was done last time, for the environment might have changed. We aim to teach robots daily manipulation tasks using human demonstrations so that they are able to fulfill them in a changing environment. To learn how humans finish a task by manipulating an object and interact with the environment, we need 3-dimensional motion data of the objects involved in fine manipulation motion, and data that represent the interaction.

Most of the currently available motion data are in the form of vision, i.e., RGB videos and depth sequences (for example, Fathi et al. [2012], Rohrbach et al. [2012], Shimada et al. [2013], Das et al. [2013], Kuehne et al. [2014], Fathi et al. [2011], Rogez et al. [2014]), which are of little or no direct use to our purpose. Nevertheless, certain datasets exist which do include motion data. Slice & Dice dataset Pham and Olivier [2009] includes 3-axis acceleration of cooking utensils which are used while salads and sandwiches are prepared. 50 Salad dataset Stein and McKenna [2013] includes 3-axis acceleration of more cooking utensils than Slice & Dice which are involved in salad preparation. CMU-MMAC de la Torre et al. [2009] dataset includes motion capture and 6-degree of freedom (DoF) inertia measurement unit (IMU) data of the human subjects while the subjects are making dishes. The IMUs record acceleration in (x, y, z, yaw, pitch, roll). The Actions-of-Making-Cereal Pieropan et al. [2014] dataset includes 6-DoF pose trajectories of the objects involved in cereal making that are estimated from RGB-D videos. The TUM Kitchen dataset Tenorth et al. [2009] includes motion capture data of the human subjects while the subjects are making dishes. The OPPORTUNITY dataset Roggen et al. [2010] includes 3-D acceleration and 2-D rotational velocity of objects. The Wrist-Worn-Accelerometer Bruno et al. [2014] dataset includes 3-axis acceleration of the wrist while the subject is doing daily activities. The Kinodynamic dataset Pham et al. [2016] includes mass, inertia, linear and angular acceleration, angular velocity, and orientation of the objects, but the manipulation exists in its own rights and does not serve to finish a task.

The aforementioned datasets are less than ideal in that 1) calculating the position trajectory using the acceleration may be inaccurate due to accumulated error, 2) the motions of objects are not always emphasized or even available, and 3) all the activities are not fine manipulations that serve to finish tasks. Having identified those deficiencies, we collected a dataset ourselves that includes 3-dimensional “position and orientation, force and torque” data of tools/objects being manipulated that fulfill certain tasks. The dataset is potentially suitable for learning either motion Huang and Sun [2015] or force Lin et al. [2012] from demonstration, motion recognition Subramani et al. [2017]; Aronson et al. [2016] and understanding Aksoy et al. [2011]; Paulius et al. [2018]; Flanagan et al. [2006]; Soechting and Flanders [2008], and is potentially beneficial to grasp research Lin and Sun [2015a,b]; Sun et al. [2016].

2 Overview
We present a dataset of daily interactive manipulation. Specifically, we record daily performed fine motion in which an object is manipulated to interact with another object. We refer to the person who executes the motion as subject, the manipulated object as tool, and the interactive object as object. We focus on recording the motion of the tool. In some cases, we also record the motion of the object.

The authors are with the department of Computer Science and Engineering, University of South Florida, Tampa, USA.

Corresponding author:
Yu Sun, Computer Science and Engineering, University of South Florida, 4202 E Fowler Ave., ENB 331, Tampa, FL 33620
Email: yusun@cse.usf.edu
Table 1. The count for each modality for each motion. Each motion is coded as $m_x$, where $x$ is an integer.

| Code | Name                               | Total | PO | FT | vision |
|------|------------------------------------|-------|----|----|--------|
| m2   | stir with spatula                  | 25    | 25 | 25 | 25     |
| m3   | sprinkle, shake pepper             | 40    | 40 | 40 |        |
| m4   | spread/oil                         | 25    | 25 | 25 | 25     |
| m6   | vertical cut                       | 25    | 25 | 25 | 25     |
| m7   | use spoon to pick up               | 98    | 98 | 98 | 35     |
| m8   | pizza wheel                        | 25    | 25 | 15 | 25     |
| m10  | use black brush                    | 25    | 25 | 25 | 25     |
| m11  | spear stuff using fork             | 30    | 30 | 30 | 30     |
| m12  | stir water using spoon             | 25    | 25 | 25 | 25     |
| m13  | fasten screw with screwdriver      | 40    | 40 | 40 | 40     |
| m14  | loosen screw with screwdriver      | 35    | 35 | 35 |        |
| m15  | unlock lock with key               | 165   | 165| 165| 75     |
| m16  | fasten nut with wrench             | 40    | 40 | 40 | 15     |
| m17  | use paint brush to dip and spread  | 25    | 25 | 25 | 25     |
| m18  | use hammer to hammer in nail       | 25    | 25 | 25 | 25     |
| m19  | brush teeth                        | 50    | 50 | 50 | 50     |
| m20  | use file to file wooden thing      | 125   | 125| 125| 25     |
| m21  | comb hair                          | 25    | 25 | 25 | 25     |
| m22  | scrape substrate from surface      | 25    | 25 | 25 | 25     |
| m23  | peel cucumber/potato               | 30    | 30 | 30 | 30     |
| m24  | slice cucumber                     | 25    | 25 | 25 | 25     |
| m25  | flip bread                         | 124   | 124| 124| 74     |
| m26  | use spoon to scoop and pour        | 25    | 25 | 25 | 25     |
| m27  | shave object                       | 30    | 30 | 30 | 30     |
| m28  | use roller to roll out dough       | 30    | 30 | 30 | 30     |
| m30  | loosen nut with wrench             | 46    | 46 | 46 |        |
| m31  | scoop and pour with measuring spoon/cup | 30 | 30 | 30 | 30 |
| m32  | insert pef into pegboard           | 140   | 140| 140| 10     |
| m33  | brush powder across grey tray      | 80    | 80 | 80 |        |
| m34  | insert straw through to-go cup lid | 25    | 25 | 25 |        |
| m35  | m34 with eyes closed               | 25    | 25 | 25 | 25     |
| m36  | m31 without pour                   | 120   | 120| 120|       |

The dataset consists of two parts. The first part contains 1,593 trials that cover 32 types of motions. We choose fine motions that people commonly perform in daily life which involve interaction with a variety of objects. We reviewed existing motion-related datasets [Huang et al. (2016); Huang and Sun (2016); Bianchi et al. (2016)] to help us decide which motions to collect.

The second part contains the pouring motion alone. We collect it to help with motion generalization to different environments. We chose pouring because 1) pouring is found to be the second frequently executed motion in cooking, right after pick-and-place [Paulius et al. (2016)] and 2) we can vary the environment setup of the pouring motion easily by switching different materials, cups, and containers. The pouring data contain 1,596 trials of pouring 3 materials from 6 cups into 10 containers.

We collect the two parts of the data using the same system. We specifically describe the pouring data in Sec. 10.

The dataset aims to provide position and orientation (PO) and force and torque (FT), nevertheless, it also provides RGB and depth vision with a smaller coverage. Table 1 shows the number of trials and the counts of each modality for each motion. The minimum number of trials for each motion is 25. Table 2 shows the coverage of each modality throughout the entire data, where the coverage has a range of (0, 1], and a coverage of 1 means the modality is available for every trial. The lower coverage of the vision modality is due to filming permission restriction.

| Modality | PO | FT | vision |
|----------|----|----|--------|
| Coverage | 1.0| 1.0| 0.50   |

3 Hardware

On a desk surface, we use blue masking take to enclose a rectangular area which we refer to as the working area, and within which we perform all the motions. We make a PrimeSense RGB+depth camera aim at the working area from above.

We started collecting PO data using the OptiTrack motion capture (mocap) system and soon afterwards replaced OptiTrack with the Patriot mocap system. Both systems provide 3-dimensional PO data regardless of their difference in technology. Patriot includes a source and a sensor. The source provides the reference frame, with respect to which the PO of the sensor is calculated. We use an ATI Mini40 force and torque (FT) sensor together with the Patriot PO sensor. To attach both the FT sensor and the PO sensor to a tool, we use a cascading structure that can be represented as: (tooltip + adapter + FT sensor + universal handle + PO sensor), where “+” means “connect”. The end result is shown in Fig. 1. A tool in general consists of a tooltip and a handle. We disconnect the tooltip from the stock handle, insert the tooltip into a 3D-printed adapter, and glue them together. Then we connect the adapter with the tooling side of the FT sensor using screws. We 3D-print a universal handle and connect it with the mounting side of the FT sensor using screws. At the end of the universal handle we mount the PO sensor using screws. In some cases, we track the object in addition to the tool, and to do that we put a second PO sensor on the object, as shown in Fig. 2.
3

Figure 3. Examples of the tools that we have adapted

Figure 4. The base point of the tool is the center of the tooling side of the FT sensor

Each tooltip is provided with a separate adapter. Since the tooltip and the adapter is glued together, a tool is equivalent to “tooltip + adapter”. Fig. 3 shows the tools that we have adapted.

4 Coordinate frames

To track a tool using OptiTrack, we need to define the ground plane and define the tool as a trackable. The ground plane is set by aligning a right-angle set tool to the bottom left corner of the working area. The trackable is defined from a set of selected markers, and is assigned the same coordinate frame, with the origin being the centroid of the markers. This is shown in Fig. 5.

Patriot contains a source that supports up to two sensors. The source provides the reference frame for the sensors as shown in Fig. 6. We define the base point of the tool to be the center of the tooling side of the FT sensor, as shown in Fig. 4. The translation from the PO sensor to the base point of the tool is \([14.3, 0, 0.7]\), in the frame of the PO sensor, unit centimeter.

The FT sensor and the PO sensor are connected through the universal handle. The groove on the universal handle is orthogonal to both the \(x - y\) plane of the FT sensor and the \(y - z\) plane of the PO sensor. The relationship between the local frames of the FT sensor and the PO sensor is shown in Fig. 7.

5 Calibrate FT

Definition 1. The level pose of the universal handle is a pose in which the groove of the handle faces up, and in which the \(y - z\) plane of the FT sensor or equivalently the \(x - y\) plane of the PO sensor is parallel to the desk surface.

Definition 2. An average sample is the average of 500 FT samples.

The FT sensor has non-zero readings when it is static with the tool installed on it. We calibrate the FT sensor, or make the readings zeros, before we collect any data. We hold the handle in a level pose (Definition 1), and take an average sample (Definition 2) which we set as the bias \(F_T\). We subtract the bias from each FT sample before saving the
sample: \( FT_t \leftarrow FT_t - FT_b \). We calibrate the FT sensor each time we switch to a new tool.

### 6 Modality Synchronization

Different modalities run at different frequencies and therefore need synchronization, which we achieve by using time stamps. We use Microsoft QueryPerformanceCounter (QPC) to query time stamps with millisecond precision.

When we start the collection system, we query the time stamp and set it as the global start time \( t_0 \). Then we start each modality as an independent thread, so that they run simultaneously and do not affect each other. For each sample, a modality queries the time stamp \( t \) through QPC, and set the difference between \( t \) and \( t_0 \), i.e. the elapsed time since \( t_0 \) as the time stamp for that sample:

\[
t ← t − t_0.
\]  

### 7 Data Format

The data are organized in a “motion → subject → trial → data files” hierarchy, as shown in Fig. 8 where the prefixes for motion, subject, and trial directories are \( m \), \( s \), and \( t \), respectively.

RGB videos save as .avi, depth images save as .png, and the rest data files save as .csv. Both RGB and depth have a resolution of 640×480, and are collected at 30Hz.

The .csv files excluding those of OptiTrack follow the same structure as shown in Fig. 8. The first row contains the global start time and is the same in all the .csv files that belong to the same trial. Starting with the second row, each row is a data sample, of which the first column is the time stamp \( t \) through QPC, and set the difference between \( t \) and \( t_0 \), i.e. the elapsed time since \( t_0 \) as the time stamp for that sample:

\[
t ← t − t_0.
\]  

### 8 Using the data

We provide MATLAB code that visualizes the PO data for OptiTrack as well as Patriot, as shown in Fig. 12. The visualizer displays the trail of the base point of the tool (Fig. 4) and the object if applicable as the motion is played as an
Figure 12. Visualizing the PO data

animation in 3D. The user can also manually slide through the motion forward or backward and go to a particular frame.

The FT and PO csv files have multiple formats, and we provide Python code that extracts FT and PO data from each trial given the path of the root folder. Although we have explained the format of the csv files of the FT and PO data in Sec. 7, we highly recommend using our code to get the FT and PO data to avoid error.

Each modality is sampled at a unique frequency, and using multiple modalities requires using the time stamps. One or more modalities need upsampling or downsampling.

9 Known issue

The PO data recorded using OptiTrack contain occasional flickering and stagnant frames. This is caused by the dependency of OptiTrack on the line of sight. This issue is not present in the data collected with Patriot.

10 The pouring data

We want to learn to perform a type of motion from its PO and FT data, and generalize it, i.e., execute it in a different environment. Thus, we need data that show how the motion vary in multiple different environments. We realize that since pouring is the second frequently executed motion in cooking [Paulius et al. 2016], it is worth learning. Also, collecting pouring data that contain different environment setup is easy thanks to the convenience of switching material, cups, and containers. Therefore, we collected the pouring data.

The pouring data include FT, Patriot PO, and RGB videos (no depth). We collected the data using the same system as described above. In the following, we explain what has not been covered and what differs from above.

The physical entities involved in a pouring motion include the material to be poured, the container from which the material is poured which we refer to as cup, and the container to which the material is poured which we refer to as container. The pouring data contain 1,596 trials of pouring water, ice, and beans from six different cups to ten different containers. Cups are considered as tools and are installed on the FT sensor through 3D-printed adapters.

A second PO sensor is taped on the outer surface of the container just below the mouth.

We collect the FT data differently from above. When the cup is empty, we hold the handle in a level pose (Definition 1), and take an average sample (Definition 2) which we call "FT_empty". Then we fill the cup with the material to an amount we desire, hold the handle in a level pose, and take an average sample which we call "FT_init", Then we pour, during which we take however many samples (not average samples) which we call "FT". After we finish pouring, we hold the handle in a level pose, and take an average sample which we call "FT_final". In summary, we save four kinds of FT data files – three contain an average sample each: FT_empty, the FT_init, FT_final, and one contains regular samples: FT. We do not consider bias.

The organization of the data is shown in Fig. 13.

The pouring data can be used to learn how to pour in response to the sensed force of the cup. The force is a non-linear function of the physical properties of the cup and the material, the speed of pouring, the current pouring angle, the amount of remaining material in the cup, as well as other possibly related physical quantities, [Huang and Sun 2017] shows an example of modeling such function using recurrent neural network and generalizing the pouring skills to unseen cups and containers.

11 Conclusion & Future work

We have presented a dataset of daily interactive manipulation. The dataset includes 32 types of motions, and provides position and orientation, and force and torque for every motion trial. In addition, to support motion generalization to different environments, we chose the pouring motion and collected corresponding data. We plan to extend the collection to other types of motions in the future.

12 Acknowledgments

This material is based upon work supported by the National Science Foundation under Grants No. 1421418 and No. 1560761.

References

Aksoy EE, Abramov A, Drr J, Ning K, Dellen B and Wegter F (2011) Learning the semantics of objectaction relations by observation. The International Journal of Robotics Research 30(10): 1229–1249.

Aronson R, Bhatia A, Jia Z, Guillame-Bert M, Bourne D, Dubrawski AW and Mason MT (2016) Data-driven classification of screwdriving operations. In: International Symposium on Experimental Robotics.
Bianchi M, Bohg J and Sun Y (2016) Latest datasets and technologies presented in the workshop on grasping and manipulation datasets. arXiv preprint arXiv:1609.02531.

Bruno B, Mastrogiovanni F and Sgorbissa A (2014) A public domain dataset for all recognition using wrist-placed accelerometers. In: Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on. pp. 738–743.

das P, Xu C, Doell R and Corso J (2013) 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.

de la Torre F, Hodgins J, Bargteile A, Collado A, Martin X, Macey J and Beltran P (2009) Guide to the Carnegie Mellon University multimodal activity (cmu-mm) database. Technical Report CMU-RI-TR-08-22, Robotics Institute, Carnegie Mellon University.

Fathi A, Li Y and Rehg JM (2012) Learning to recognize daily actions using gaze. In: Proceedings of the 12th European Conference on Computer Vision - Volume Part I, ECCV’12. pp. 314–327.

Fathi A, Ren X and Rehg JM (2011) Learning to recognize objects in egocentric activities. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. pp. 3281–3288. DOI:10.1109/CVPR.2011.5995444.

Flanagan JR, Bowman MC and Johansson RS (2006) Control strategies in object manipulation tasks. Current Opinion in Neurobiology 16(6): 650–659. Motor systems / Neurobiology of behaviour.

Huang Y, Bianchi M, Liarokapis M and Yu S (2016) Recent data sets on object manipulation: A survey. Big Data 4(4): 197–216.

Huang Y and Sun Y (2015) Generating manipulation trajectory using motion harmonics. In: Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on. IEEE, pp. 4949–4954.

Huang Y and Sun Y (2016)Datasets on object manipulation and interaction: a survey. arXiv preprint arXiv:1607.00442.

Huang Y and Sun Y (2017) Learning to pour. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE. Accepted.

Kuehne H, Arslan A and Serre T (2014) 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.

Lin Y, Ren S, Clevenger M and Sun Y (2012) Learning grasping force from demonstration. In: Robotics and Automation (ICRA), 2012 IEEE International Conference on. IEEE, pp. 1526–1531.

Lin Y and Sun Y (2015a)Grasp planning to maximize task coverage. The International Journal of Robotics Research 34(9): 1195–1210.

Lin Y and Sun Y (2015b) Task-based grasp quality measures for grasp synthesis. In: Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on. IEEE, pp. 485–490.

Lin Y and Sun Y (2016) Task-oriented grasp planning based on disturbance distribution. In: Robotics Research. Springer, pp. 577–592.

Paulius D, Huang Y, Milton R, Buchanan WD, Sam J and Sun Y (2016) Functional object-oriented network for manipulation learning. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 2655–2662.

Paulius D, Jelodor A and Sun Y (2018) Functional object-oriented network: Construction & expansion. In: Icra. IEEE, pp. 1–6.

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

Pham TH, Kyriazis N, Argyros AA and Kheddar A (2016) Hand-Object Contact Force Estimation From Markerless Visual Tracking. Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence.

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

Rogez G, III JSS, Khademi M, Montiel JMM and Ramanan D (2014) 3D hand pose detection in egocentric RGB-D images. CoRR abs/1412.0065.

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

Rohrbach M, Amin S, Andriluka M and Schiele B (2012) A database for fine grained activity detection of cooking activities. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on.

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

Soechting JF and Flanders M (2008) Sensorimotor control of contact force. Current Opinion in Neurobiology 18(6): 565 – 572.

Stein S and McKenna SJ (2013) 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. pp. 729–738.

Subramani G, Rakita D, Wang H, Black J, Zinn M and Gleicher M (2017) Recognizing actions during tactile manipulations through force sensing. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE. Accepted.

Sun Y, Lin Y and Huang Y (2016) Robotic grasping for instrument manipulations. In: Ubiquitous Robots and Ambient Intelligence (URAi), 2016 13th International Conference on. IEEE, pp. 302–304.

Tenerth M, Bandouch J and Beetz M (2009) The turn 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. pp. 1089–1096.