Data paper

The International Journal of
Robotics Research
2015, Vol. 34(3) 251–255

DOI: 10.1177/0278364914555720
ijr.sagepub.com
SAGE

The Yale human grasping dataset: Grasp, object, and task data in household and machine shop environments

Ian M. Bullock, Thomas Feix and Aaron M. Dollar

Abstract

This paper presents a dataset of human grasping behavior in unstructured environments. Wide-angle head-mounted camera video was recorded from two housekeepers and two machinists during their regular work activities, and the grasp types, objects, and tasks were analyzed and coded by study staff. The full dataset contains 27.7 hours of tagged video and represents a wide range of manipulative behaviors spanning much of the typical human hand usage. We provide the original videos, a spreadsheet including the tagged grasp type, object, and task parameters, time information for each successive grasp, and video screenshots for each instance. Example code is provided for MATLAB and R, demonstrating how to load in the dataset and produce simple plots.

Keywords

Grasping, manipulation, multifingered hands, dexterous

1. Synopsis

We provide a large annotated video dataset of housekeeper and machinist grasping in unstructured environments. A head-mounted camera is used to record the hands and their interaction with the environment. For each instance of grasp in the video (right hand only), the data is tagged with grasp type, properties of the object including size, shape, stiffness, and mass parameters, and task properties including force, movement constraints, and general class parameters (Figure 1). The dataset was used in previous publications to analyze human grasp usage, and the interaction between grasp choice and the object and task properties (Bullock et al., 2013; Feix et al., 2014a, 2014b). The full dataset, with raw video and the tagged data, can be downloaded at http://www.eng.yale.edu/grablab/humangrasping/. Note that the Massachusetts Institute of Technology (MIT) license is used for the example code, and the .csv data is available under a Creative Commons Attribution 4.0 International (CC BY 4.0) license (Creative Commons Corporation, 2014). However, the authors maintain copyright for the video and image data, with download and use permission granted for research use only. Permission of the authors should be obtained prior to distribution of video or image data, including modified versions. The authors want free use of the video and images for any research purposes, but the video and images should not be redistributed for any other purpose.

2. Methods

2.1. Participants

Two machinists and two housekeepers were recorded. "Machinist 1" is a 41-year-old male with more than 20 years of professional machining experience, and "Machinist 2" is a 50-year-old male with about 30 years of experience. "Housekeeper 1" is a 30-year-old female with one year of housekeeping experience, and "Housekeeper 2" is a 20-year-old female with eight months of experience. All subjects have normal physical ability, are right handed, and were able to generate at least 8 hours of data.

2.2. Experimental procedure and apparatus

The participants wore the head-mounted camera shown in Figure 2 during their normal work. A total of at least 8 hours of hand usage was recorded for each subject, over multiple days. The participants confirmed that the video

Yale GRAB Laboratory, Department of Mechanical Engineering & Materials Science, Yale University, USA

Corresponding author:
Ian M. Bullock, Department of Mechanical Engineering & Materials Science, Yale University, New Haven, CT 06511, USA.
Email: ian.bullock@yale.edu

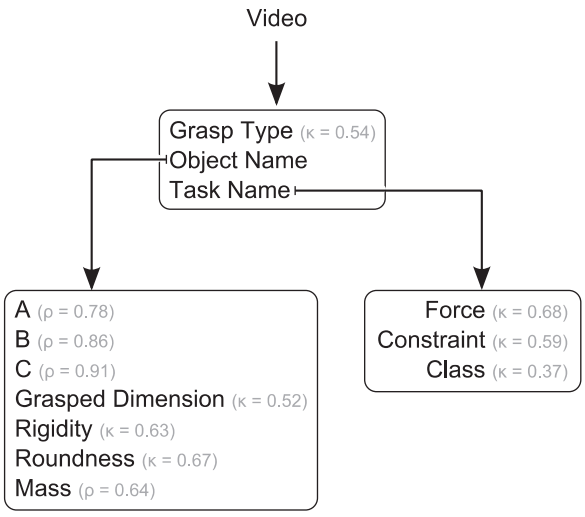

Fig. 1. Overview of the data. First, the grasp type and the high-level object and task names were assigned. Object and task properties were then added based on the object and task names. Cohen's κ and Pearson correlation ρ give an estimate of the achieved inter-rater reliability.

recorded was representative of the general set of tasks that they perform for their profession.

The hardware consists of a tube camera (RageCams, model 3225, 200 g, 22 mm dia × 60 mm long, 640 × 480 resolution) with a wide-angle fisheye lens (2.5 mm, ~140° field of view) attached to a three-band head strap. The camera is connected to a mini digital video recorder (AngelEye 2.4 GHz PVR, 115 mm × 65 mm × 25 mm, 25 FPS). The original recorder broke partway through the study and was replaced with as close a model as possible. The video from the newer model can be identified by the presence of yellow timestamp text (rather than white). An external battery pack (12 V) powers the camera. The overhead view, similar to that used by Kemp (2005), was chosen because it shows the entire workspace of both arms in front of the body as well as enough of the surroundings to give the context of the grasps. Figure 2 shows two sample images taken with this setup.

2.3. Data annotation

The annotation was done in two stages (Figure 1). In the first stage, the grasp type and the high-level task and object names were recorded. The full set of grasps used are those by Feix et al. (2009), but the original names from Cutkosky (1989) are used when possible. For the second annotation stage, each object and task name was assigned a number of additional properties according to the classification schemes fully described by Feix et al. (2014a, 2014b).

For the first stage, also described by Bullock et al. (2013), two researchers trained in classifying grasps monitored the slowed-down video. The raters came from an engineering background and all were familiar with human grasping literature. They were given formal rating guidelines, as well as a "cheat sheet" showing visually all the grasp types and their names. The coding guidelines were such that whenever the subject changes their grasp, acquires an object, or releases an object, the new grasp state is recorded, along with the timestamp at which the switch was made. Quick grasp transitions lasting less than a second are not recorded. In addition, the object that the subject grasps and a description of the task performed are recorded. Only data for the right (dominant) hand is recorded. In cases of occlusion, the continuous nature of the video generally allowed the raters to guess the grasp with a high degree of certainty. In extreme cases, the raters did occasionally mark grasps as "unknown."

Each video segment was tagged by one of the two researchers. A single rater per segment was used in order to allow much more video data to be analyzed in a reasonable timeframe, as well as due to the extensive training required for each rater. Since the original video is included, further tagging could be added in the future as desired.

After the initial grasp/object/task tagging, the second stage of tagging involved assigning further properties based on the object and task taxonomies described by Feix et al. (2014a, 2014b). Specifically, two raters assigned seven object properties to each object name, and three task properties to each task name, based on both the name itself and a group of video snapshots associated with that object or task description. Generally, the amount of variation within

Fig. 2. Camera setup and example images. A head-mounted camera (left) was used to record two machinists and two housekeepers. A sample video image is shown for one machinist and housekeeper participant.

Table 1. Common data subsets used in previous publications. The pseudocode conditions indicate how these subsets can be obtained from the full dataset, to facilitate comparison of results to the existing work.

	Instances	Publication	Condition pseudocode
Full dataset	18,210	(Bullock et al., 2013)	-
Grasp present	11,539	(Bullock et al., 2013)	Grasp != "no grasp"
Grasp & object present	9100	(Feix et al., 2014b)	Grasp != "no grasp" & CCObj == false
Grasp & task present	9933	(Feix et al., 2014a)	Grasp != "no grasp" & CCTask == false
Grasp, object, & task present	7770	(Feix et al., 2014a)	Grasp != "no grasp" & CCObj == false & CCTask == false

a given object or task description was small. In some cases, if the snapshots indicated the object or task description was too broad, the raters instead tagged it as "cannot classify" (CC). If either of the two raters decided that classification is not possible, that object or task was not used in the further analysis in Feix et al. (2014a, 2014b). After the two raters assigned their ratings, one rater was given the final say in deciding which rating to keep in cases of disagreement. This step added a final review of the data to help reduce any errors from either rater.

A small amount of data cleanup was required after the tagging process. A cable reliability issue caused the video to go black during a small proportion of the housekeeper 1 data, reducing the data duration to 7.45 hours. The data from the other subjects were then trimmed down to the 7.45 hour duration from housekeeper 1 to match the subject data length. The other main cleanup step was to handle a few instances where multiple grasps were recorded by the raters, usually when the subject was carrying multiple objects with their dominant hand. For these instances, the principal first grasp is taken.

Between the studies of Bullock et al. (2013) and Feix et al. (2014a, 2014b), some additional cleanup was performed. Specifically, some damaged video files could not be used and were removed from the dataset, making the final data duration 6.9 hours per participant. Note that the overall grasp frequencies all changed by less than 1% as a result of this change, showing that the slight reduction in data duration should have little impact on the results. After this additional cleanup, the final dataset includes 27.7 hours of data.

2.4. Inter-rater agreement

Since two raters were used to analyze the video used in this study (approximately 50% of the data per rater), an inter-rater reliability assessment was performed using a modified Cohen's κ method (Cohen, 1960). Since the data does not involve discrete "questions," the confusion matrix was created by recording the durations of agreement or disagreement in the tagged grasp over the same sample of data, as suggested by Conger (1985). Two 1-hour samples of data were prepared from several different videos, 1 hour from the machinists and 1 hour from the housekeepers. While the samples were mainly taken from two of the subjects (housekeeper 1 and machinist 1), the types of grasps in the sample set should still be representative of the four subjects, since very similar tasks were being performed by the pairs of subjects in each profession.

The 1-hour housekeeper sample was rated at the beginning of the study, while the machinist sample was rated after completion of the study. Thus, the housekeeper sample can be seen as a best case view of the rater reliability, while the machinist sample is a worst case view, since ratings can drift over the course of a study. Because of this, we have opted to average the two samples to produce an overall confusion matrix. The full confusion matrix is available in the *confusionMatrixTotal.csv* file. Cohen's κ was calculated using this confusion matrix, giving $\kappa = 0.54$. This represents the proportion of agreement that is not due to chance. The value results from various types of errors, including timing discrepancies and difficult to distinguish grasps. For a full discussion of the grasp inter-rater data, please see Bullock et al. (2013).

Inter-rater assessment was also used for each of the grasp and task properties. The final Pearson correlation (ρ) and Cohen's κ for these properties can be seen in Figure 1. Some values are particularly high, such as the correlation for the major object dimensions ($\rho = 0.8$–0.9), while other properties, such as the task class, proved much harder for human raters to classify consistently ($\kappa = 0.37$). Overall, the inter-rater can be used to estimate uncertainty of future results, as well as to help better understand which descriptions of grasp, task, and object data are most clearly defined, and which ones could be improved through future classification work.

3. Dataset structure and usage

The full dataset consists of 18,210 grasp instances. Depending on whether task and object data are required, the number of instances is reduced further. Table 1 gives an overview of the subsets of the data used in previous publications. For example, if task and grasp data are required, there are 9933 instances that meet this condition (Grasp != "no grasp" & CCTask == 0). The dataset parameters are summarized in Table 2.

In addition to the dataset, also the video files on which it is based are provided. Due to privacy concerns, all frames containing faces or other private information were blacked out. This included, for example, cell phones, mail, family pictures, and calendars. Overall this step blacked out 8.2% of the video. Re-encoding of the video during this stage

Table 2. Overview of all fields in the tagged dataset.

Parameter	Description	Data	
Video	Number of the video file	Video number from 1–179	
Time stamp	Time stamp of the grasp in the video file	Video timestamp in hh:mm:ss format	
Duration	Length of the grasp instance	Duration in seconds	
Subject	Participant profession and number	Machinist 1/2, Housekeeper 1/2	
BlackRatio	Proportion of instance blacked out for privacy	Ratio between 0 (all visible) and 1 (all black)	
Grasp	The grasp type according to (Feix et al., 2009)	no grasp, one of 33 grasp types	Grasp
OppType	Opposition type of the grasp (Mackenzie & Iberall, 1994)	Pad, Palm, Side, NG	
PIP	Power, intermediate or precision grasp	Power, Intermediate, Precision, NG	
Object	High-level object name	No object, object name	Object
A	Longest object dimension	Length in cm	
B	Intermediate object dimension	Length in cm	
C	Shortest object dimension	Length in cm	
Grasped dimension	Dimension along which object is grasped	a/b, a/b/c, b, c, b/c, floppy, CCObj, NG	
Rigidity	Rigidity of the object	rigid, fragile, squeezable, floppy, CCObj, NG	
Roundness	Dimensions along which object is round	a, abc, c, non-round, floppy, CCObj, NG	
Mass	Mass of the object	Value in g	
CCObj	True (1) if Cannot Classify Object (see Section 3)	0, 1, NG	
Shape	Basic shape class, according to Zingg (1935)	equant, oblate, prolate, bladed, CCObj, NG	
Type	Object type, as defined by Feix et al (2014b)	11 object types, CCObj, NG	
Task	High-level task name	no task, brief task description	Task
Force	Type of forces required for task	weight, interaction, CCTask, NG	
Constraint	Constraints of the task	11 constraint types, CCTask, NG	
Class	Function class of the task	hold, feel, use, CCTask, NG	
CCTask	True (1) if Cannot Classify Task (see Section 3)	0, 1, NG	
		NG = No Grasp, CCTask/CCObj = Cannot Classify	

was performed. Parameters were manually adjusted to reduce file size as much as possible while not significantly reducing video quality, according to qualitative inspection. MPEG-4 codec was used with a quality setting of 60 in MATLAB, resulting in a bitrate of about 2000 kbps. The column *BlackRatio* in the dataset indicates the ratio of a particular sample that has been blacked out in the video, from 0 to 1, where 0 would indicate no blacking out of that sample, and 1 would indicate the sample has been completely removed in the video. A small number (0.02%) of the original video frames were found to be corrupt and were also blacked out.

To facilitate quick usage of the data, examples are provided in the MATLAB and R programming languages, but the main dataset is in a simple comma separated value (csv) file that should be easy to load in any language. These examples, in files *demoScript.m* and *demoScript.R*, show how to load the data in and produce some simple plots and calculations from the data. The confusion matrix provided in *confusionMatrixTotal.csv* can be used with statistical simulation methods to help estimate the uncertainty present in future calculations, as in Bullock et al. (2013). This data can also provide insight into which grasp descriptions may be similar or interchangeable, for future development of grasp analysis techniques.

Acknowledgements

The authors would like to thank Charlotte Guertler, Joshua Zheng, and Sara De La Rosa for their work in coding the video dataset used in this paper.

Funding

This work was supported in part by the National Science Foundation (grant NSF IIS- 0953856).

References

Bullock IM, et al. (2013) Grasp frequency and usage in daily household and machine shop tasks. *IEEE Transactions on Haptics* 6(3): 296–308.

Cohen J (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1): 37–46.

Conger AJ (1985) Kappa reliabilities for continuous behaviors and events. *Educational and Psychological Measurement* 45(4): 861–868.

Creative Commons Corporation (2014) Creative Commons - Attribution 4.0 International - CC BY 4.0. Available at: http://creativecommons.org/licenses/by/4.0/legalcode (accessed 10 July 2014).

Cutkosky MR (1989) On grasp choice, grasp models, and the design of hands for manufacturing tasks. *IEEE Transactions on Robotics and Automation* 5(3): 269–279.

Feix T, et al. (2009) A comprehensive grasp taxonomy. In: *robotics, science and systems: workshop on understanding the human hand for advancing robotic manipulation*, Seattle, WA.

Feix T, Bullock IM and Dollar AM (2014a) Analysis of human grasping behavior: Correlating tasks, objects and grasps. *IEEE Transactions on Haptics*. Epub ahead of print 24 September 2014. DOI: 10.1109/TOH.2014.2326867.

Feix T, Bullock IM and Dollar AM (2014b) Analysis of human grasping behavior: Object characteristics and grasp type. *IEEE Transactions on Haptics* 7(3): 311–323.

Kemp CC (2005) *A wearable system that learns a kinematic model and finds structure in everyday manipulation by using absolute orientation sensors and a camera*. Massachusetts Institute of Technology.

Mackenzie CL and Iberall T (1994) *The Grasping Hand, Volume 104 (Advances in Psychology)*. 1st ed. Amsterdam: North Holland.

Zingg T (1935) *Beitrag zur Schotteranalyse*. PhD Thesis, ETH Zürich, Switzerland.