2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

# On Human Grasping and Manipulation in Kitchens: Automated Annotation, Insights, and Metrics for Effective Data Collection

Nathan Elangovan[1,*], Ricardo V. Godoy[1,*], Felipe Sanches[1], Ke Wang[2], Tom White[2], Patrick Jarvis[2], and Minas Liarokapis[1]

***Abstract*— The advancement in robotic grasping and manipulation has elicited an increased research interest in the development of household robots capable of performing a plethora of complex tasks. These advancements require the shift of robotics research from a laboratory setting to dynamic and unstructured home environments. In this work, we focus on a comprehensive data collection and analysis of key attributes involved in the selection of grasping and manipulation strategies for the successful execution of kitchen tasks. An unprecedented dataset that comprises over 7 hours of high-definition videos that were analyzed to classify more than 10,000 kitchen activities annotated with 24 attributes each has been created. Machine learning techniques were employed to automate the annotation process partially by extracting grasp types, hand, and object information from the videos. The annotated dataset was analyzed using clustering algorithms to identify underlying patterns. This study also identifies key attributes and specific data that require focus during data collection based on inter-subject variability. The insights from this study can be used to improve the speed, quality, and effectiveness of data collection. It also helps identify the strategies employed by the humans for the execution of kitchen tasks and transfer the necessary skills to a robotic end-effector enabling it to complete the tasks autonomously or collaborate with humans.**

## I. Introduction

With the advancement in technology, a number of research studies are focusing on employing robots capable of assisting/completing activities of daily living in home environments [1], [2]. There has also been an increasing interest with respect to kitchen tasks, especially cooking [3], [4]. While existing solutions like the Moley robotics kitchen automate the process of cooking, they require a custom-designed dedicated space for all equipment and a very high cost for installation [5]. One of the key challenges associated with employing robots in a kitchen is the lack of clear demarcation of robot workspace and the highly unstructured dynamic home environment [6]. Hence, in order for efficient skill transfer to robots capable of performing in a home environment, the data needs to be captured from real homes [6], [7]. Several researchers have collected datasets in the most varied environments, from houses to machine shops [8], aiming to gather information on grasp and manipulation strategies employed by subjects for performing specific types of tasks [9]–[11]. In order to analyze large video datasets, the classic approach is to employ trained human annotators to extract the most pertinent information from the videos. However, this process requires time and human resources to be completed. Machine learning techniques can automatically analyze the videos and decrease the need for prior human knowledge of the data. Moreover, identifying patterns and clustering the data can guide future data collection and improve the dataset analysis process by making the process more efficient and accurate.

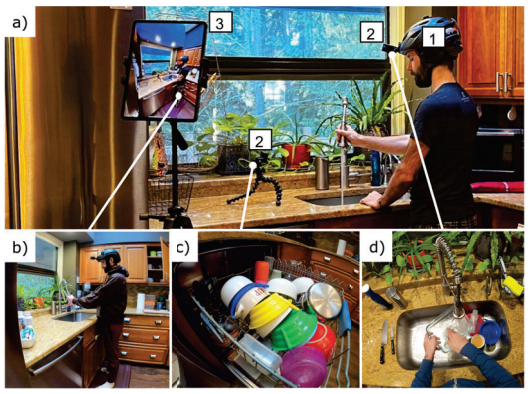


Fig. 1. a) Data collection setup at a user's home kitchen. b) Side camera view that captures the whole-body motion of the user (including shoulder and bending movements) using 3 supplementary cameras. c) Close-up side camera views (in places of occlusion, e.g. inside dishwasher or cabinets) are captured using 2 GoPro cameras. d) Egocentric view of the activities captured using 2 GoPro cameras mounted on Helmet.

In this paper, we provide a comprehensive dataset containing annotated kitchen activities created from videos of users performing the tasks in their kitchen environment. The primary focus of this data collection and analysis is to identify humans' unique grasping and manipulation strategies for successfully executing various task categories in a kitchen environment. High-definition videos of multiple users performing kitchen tasks in their home environment are captured, analyzed, and annotated with all the key attributes necessary for efficient skill transfer to robots, such as activity description, object properties, and grasping and manipulation strategies employed, among others. We then apply machine learning (ML) algorithms to identify underlying patterns in the annotated data. Employing ML techniques would enable the identification and transfer of key skills for the

[1]Nathan Elangovan, Ricardo V. Godoy, Felipe Sanches, and Minas Liarokapis are with the New Dexterity research group, Department of Mechanical and Mechatronics Engineering, The University of Auckland, New Zealand. E-mails: {sela886,rdeg264, fsan668}@aucklanduni.ac.nz, minas.liarokapis@auckland.ac.nz

[2]Ke Wang, Tom White, and Patrick Jarvis are with Acumino, USA. E-mails: {kewang,tomw,patrickj}@acumino.ai

*These authors contributed equally to this work.

development of robotic grippers and hands that can perform on par with humans as well as collaborate with humans. We provide inputs for improving data collection effectiveness, quality, and speed by automating the annotation process. The analysis also provides directions for the specific data to be collected and key attributes that require focus during data collection to improve the skill transfer and collaboration between humans and robots.

The rest of the paper is as follows: Section II presents an overview of the related work on the available kitchen environments datasets. Section III explains the data collection and annotation process and presents the framework used for automating the annotation process, Section IV shows the effectiveness of the automatic annotation and discusses the analysis and implications of the dataset, while Section V concludes the paper and discusses future directions.

## II. Related Work

The kitchen environment has been gaining attention over the last few years since different everyday tasks can be performed in this environment. This environment is constantly changing, and the objects encountered vary widely, from cutlery to appliances. While some available datasets focus on recording subjects performing a specific type o kitchen task, such as setting the table [12] or preparing a meal [13], [14], recent works are trying to capture a more significant sampling of tasks, in order to represent the dynamic nature of the kitchen environment.

The Bimanual Actions Dataset [9] contains RGB-D videos recorded from 6 subjects performing five different tasks in a kitchen context from a robot's point of view. The LEMMA Dataset [10] recorded the data in 7 different houses, performed by eight individuals in 14 kitchens/living rooms. Kitchen activities involved in intermediate meal preparation stages are also available in the MoCA dataset [15]. The EPIC-KITCHENS-100 is another relevant dataset pertaining to the annotated egocentric videos of various unstructured activities performed by individuals in their respective kitchen [11]. However, the dataset only includes egocentric videos that are annotated with activities. The EPIC-KITCHENS-100 dataset proposes an annotation pipeline using a narration approach, in which each participant narrated their action, and an object detector based on a Mask-R-CNN [16]. These datasets can be employed to extract pertinent information, for example, identifying patterns of grasps and manipulations executed by the human hands for specific tasks/objects in a kitchen environment. The general manipulation tasks and strategies have been classified in a taxonomy based on hand-centric and motion-centric attributes of the task. They have been demonstrated to apply to both humans and robots [17]. Common terminology for classifying the human hand grasp configurations has been identified based on opposition types, thumb position, finger assignments, and grasp types in terms of precision, intermediate, and power grasps executed [18]. In home and machine shop environments, it was found that subjects use five to ten specific grasps to complete 80% of the tasks [8].

## III. Methods

### A. Data collection

We analyzed the various activities performed by three subjects in their respective home kitchen environment over multiple days. Videos were collected using multiple cameras from three home kitchens using the set-up shown in Figure 1. Apart from egocentric videos, the dataset also includes supplementary videos showing object interaction in tight occluded spaces, as well as a tertiary video providing side-on views of the subject to determine shoulder movements and bending actions required to complete a task.

### B. Data Annotation

The videos were analyzed and annotated with over 24 attributes by two teams of specialized and previously oriented annotators. We aimed to annotate all the key information necessary for the completion of the tasks, including activity description, duration of activities, object properties (size, shape, orientation, etc.), grasping strategy (bimanual or single-handed, thumb position, grasp classification, etc.), and manipulation strategy (manipulation type, the direction of manipulation, etc). For the purpose of this study, we identified and classified all the activities associated with a home kitchen environment into ten task categories. The categories are grouped into four major task groups: stocking (1.Pantry, 2.Fridge), cooking (3.Breakfast, 4.Lunch, 5.Dinner, 6.Appliances), dishwashing (7.Loading, 8.Unloading), and cleaning (9.Countertops, 10.Trash). We used an existing grasp and manipulation taxonomy to describe the strategies wherever possible [18]. Moreover, we have extended the taxonomy by adding some of the unique strategies observed in a kitchen environment. The annotation guide detailing the entire annotation process is also available online.

### C. Dataset

The dataset comprises over 7 hours of high-definition videos that were analyzed to classify more than 10,000 activities annotated with 24 attributes each. The complete dataset, videos in HD quality, and a dedicated website complementing this paper can be found at:

www.newdexterity.org/kitchendataset

### D. Automated annotation

The classical manual annotation process demands human resources and time, but it is necessary for all data to be extracted as accurately as possible. In order to speed up this process and, consequently, save resources, we have developed an automatic annotation pipeline for video annotating. The purpose of this pipeline is not to completely replace the annotator, a mark that we do not consider possible at the moment with the amount of annotated data we have, but rather to help the annotator do its job more quickly, accurately, and efficiently. Our model predicts three information from the videos that are of great interest: the type of grasp performed by the user, which hand was used in the grasp, and which object was grasped. This pipeline, shown in Figure 2, will be described in detail below.

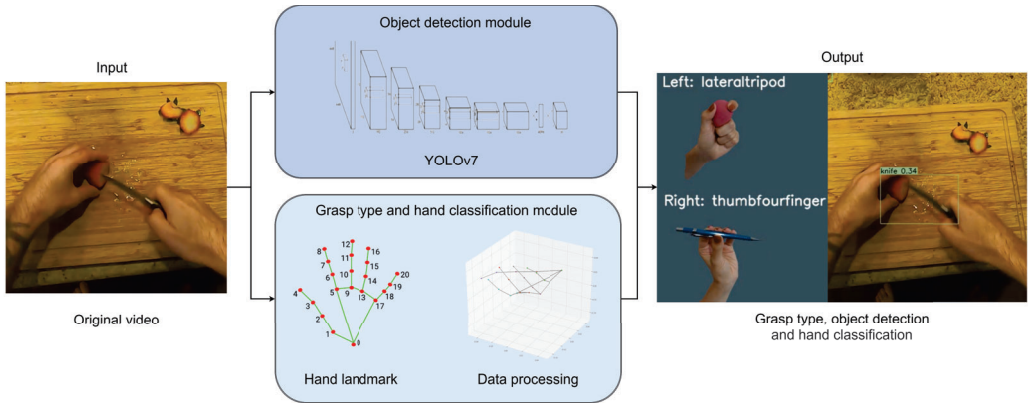


Fig. 2. Automated annotation pipeline. The recorded kitchen videos are supplied to the automated annotation pipeline, which detects the objects in the scene and classifies the hand and grasp type. This processed video with extracted information is provided to the annotator in order to make the annotation process more efficient and accurate.

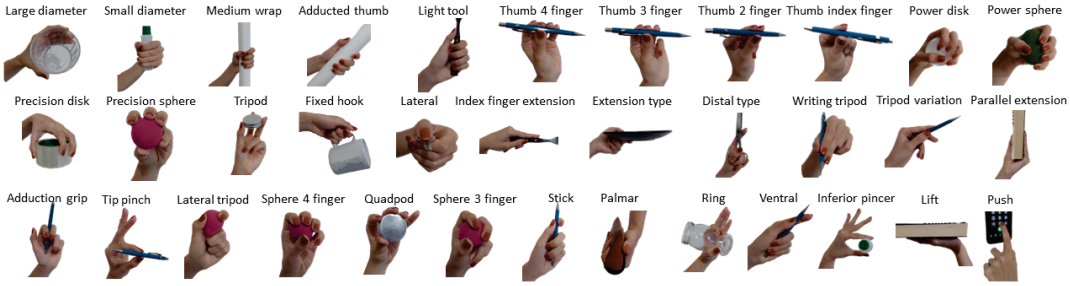


Fig. 3. Grasp taxonomy. Our grasp type classification model can predict which grasp type the user is employing to perform the task in the scene. The prediction of the model is provided to the human annotator in order to make the annotating process faster and more accurate.

*1) Grasp type classification:* In order to achieve the grasp type classification, we started by collecting 40 samples of the hand landmark of each grasp type using the Media Pipe's Hand model [19], of which 20 were collected with the user holding an object, and 20 with the user simulating the grasp without holding any object. Then, this data was preprocessed by setting the wrist as the origin and normalizing the data by the distance between the wrist and the base of the fifth digit. A Random Forest (RF) model, an ensemble method for classification, was trained in a supervised learning way to perform the grasp type classification, using the processed hand landmark data as input. The RF model discriminates between 35 grasp types, shown in Figure 3, based on the GRASP taxonomy [18]. The RF model was trained using 6-fold cross-validation, achieving an average accuracy of 79% for classifying the grasp type.

*2) Hand classification:* We employed the same Media Pipe model to collect the handedness of the detected hand.

*3) Object detection:* To perform object detection, we used the new state-of-the-art object detector YOLOv7, which recently surpassed all object detectors in accuracy and speed [20]. We employed the YOLOv7 trained on MS COCO [21]. Our automated annotation pipeline will be applied to each kitchen video collected in a post-process step before the video is provided to the professional annotators. The primary purpose of this pipeline is to give the annotator a starting point for annotations. In this way, the professional will check if the prediction is correct and, if not, correct for the one he deems correct using the model's first prediction.

### *E. Data analysis*

We employed clustering algorithms to identify underlying patterns in the multi-dimensional dataset obtained from the annotation process. The cluster analysis procedure followed in this study is presented in Figure 4. The data cleaning phase ensured the annotated dataset was continuous and converted to the required structure for processing. This included removing empty rows, identifying outliers or errors imposed during annotation, and replacing missing values in columns with NaN. Data columns that were not significant for a given analysis were also dropped.

As most of the attributes in the dataset were made up of categorical data, we employed one-hot encoding to transform them into binary values in order to be processed by the clustering algorithm. However, this resulted in a dataset with a very high dimension ($N$ columns for the $N$ unique categorical values). To overcome this limitation, the non-linear dimensionality reduction technique t-distributed Stochastic Neighbor Embedding (t-SNE) was used to reduce the dataset to two dimensions. This algorithm effectively retains local variance by retaining the variance of local points and embedding them into lower dimensions by retaining the structure of neighboring points.

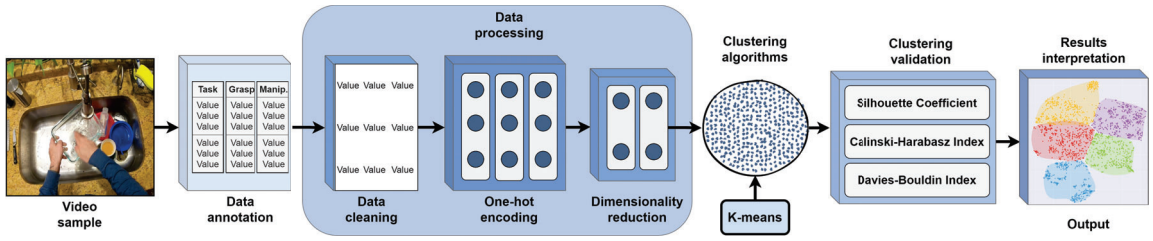


Fig. 4. The process of analyzing the dataset using clustering algorithms. The video recorded in a kitchen environment is annotated by a team of annotators. Then the data is cleaned, one-hot encoded, and have its dimensionality reduced in order to be supplied to a K-means model to cluster the dataset. The output of the clustering algorithm is validated using three different measures.

We applied K-means, one of the simplest and most popular unsupervised machine learning algorithms used in clustering to the dimensionality-reduced dataset. The optimal number of clusters "k" was calculated as 5 using the elbow method and provided as input. The outputs of the clustering algorithm were validated using three measures: the Davies-Bouldin index, Silhouette coefficient, and Calinski-Harbasz index. The higher values for silhouette's co-efficient and Calinski-Harabasz as well as the lower Davies-Bouldin index represent a higher similarity between the data points within each cluster and demonstrate the efficiency of k-means.

## IV. Results and Discussion

### A. Automated annotation

In order to evaluate the performance improvement brought by the proposed automatic annotation framework, we asked five subjects to annotate the video with and without the annotation pipeline. Each subject was asked to alternately annotate 30 seconds of video with the suggestions predicted by the framework on and then off. One video of each kitchen task category was evaluated. The process was repeated until the video came to an end. During the experiments, the time taken by each participant to annotate the set of 30 seconds of video was measured. We evaluated the average relative improvement (ARI), in percentage, by comparing how much more time the subject took to annotate the videos without the automatic annotation framework activated. Table I shows the amount of video annotated per minute of annotation, with the automatic annotation framework on and off. All subjects benefited from the suggestions predicted through the automatic annotation framework. The average relative improvement was 38.95% among the five subjects, leading to a more cost and resource-efficient annotation process.

TABLE I

Average relative improvement (ARI) for each evaluated subject when using the automatic annotation framework.

| Subject | Amount of video annotated per minute AAF On | AAF Off | ARI (%) |
|---|---|---|---|
| 1 | 0.1303 | 0.1009 | 29.23 |
| 2 | 0.0563 | 0.0477 | 18.10 |
| 3 | 0.0839 | 0.0476 | 76.22 |
| 4 | 0.2025 | 0.1347 | 50.31 |
| 5 | 0.0623 | 0.0771 | 20.90 |
| **Average** | | | **38.95** |

### B. Cluster Analysis and Interpretation

The clusters produced by the k-means algorithm had an uniform distribution with an average of around 2000 elements each, and the vital characteristics of the clusters are presented in Table II. It is evident from the observations that the points within each cluster have a high similarity and are well distinct from the points in neighboring clusters. As a majority of tasks in each cluster corresponds to one major task group, the grasping and manipulation strategy associated with these task groups can be easily identified.

For example, cluster 0 grouped activities involving the loading and unloading dishes from a dishwasher that only required four different grasps and did not require in-hand manipulation skills for completing a majority of tasks. On the other hand, cluster 2 and cluster 3 captured the cooking tasks, including the preparation of breakfast, lunch, dinner, and operating the appliances. These clusters required over 9 grasps as well as non-prehensile(objects not completely grasped/restricted) manipulation of objects to complete a significant part of the tasks. The requirement for "carry" tasks along with pick & place correlates to carrying the objects from the counter or shopping bags to the fridge/pantry during the stocking tasks, and the grasp conversion used by humans while picking and placing objects into a dishwasher can be seen from the task type category columns of cluster 4. Hence, the unsupervised learning methods employed in this study can successfully group data points into clusters enabling us to visualize the similarities and characteristics of each cluster. Furthermore, these clustering analyses show that most kitchen tasks can be successfully executed using a limited number of grasps without the need for complex in-hand manipulation tasks. A significant part of the activities performed in the kitchen environment could be classified into simple tasks like pick, place, open, close, and hold, among others that can be executed by grasping an object and moving it in space without changing the contact points.

### C. Inter-subject Variability of Strategies

The dataset was analyzed to derive specific inputs for further data collection and to determine the tasks/strategies that require more focused analysis. The percentage distribution of the various grasping and manipulation strategies employed by individual subjects for each of the ten kitchen

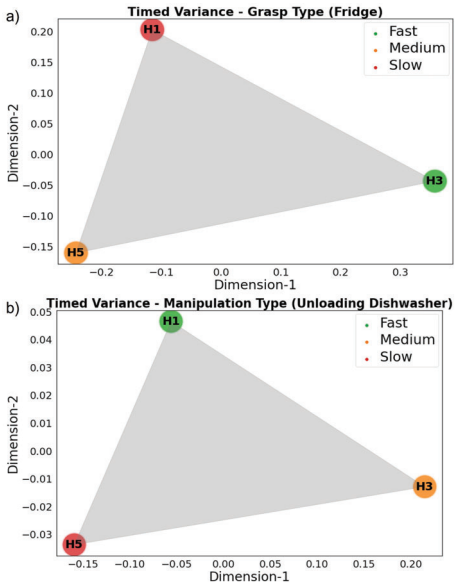


Fig. 5. The inter-subject variability mapped along with average task execution time for a) grasp type (Fridge task), and b) manipulation type (Unloading dishwasher). In such cases, the fastest task execution time would allow us to select the strategies employed by one subject over the other.

TABLE II

ANALYSIS OF THE K-MEANS CLUSTERS, INCLUDING THE DISTRIBUTION OF TASK TYPES, NUMBER OF GRASPS, MANIPULATION STRATEGIES, AND KITCHEN TASK CATEGORIES WITH MAJOR CONTRIBUTION.

| Cluster # | No. of entries | Major (>2/3rd) Contribution | | | |
|---|---|---|---|---|---|
| | | Task type | No. of Grasps | Manipul-ation type | Kitchen Task |
| 0 | 2063 | Pick, Place, Hold | 4 | MT10 | Dishwasher |
| 1 | 2456 | Pick, Place | 7 | MT10 | Cleaning (Dinner, Loading Dishwasher) |
| 2 | 1450 | Open, Close | 9 | MT10 | Cooking |
| 3 | 1930 | Pick, Place | 10 | MT10, MT8 | Cooking |
| 4 | 2120 | Hold, Pick, Place, Carry, Convert grasp | 6 | MT10, MT6 | Stocking (Loading Dishwasher, Dinner) |

tasks was calculated from the dataset. This provided us with the percent usage of over 45 different grasp types and over 15 manipulation types employed by the subjects for the successful task completion stored in a 45 dimension and 15 dimension space, respectively. The distribution of the grasp classes (power, intermediate, precision) was also calculated for the task categories. To visualize this data and enable easy computation, principal component analysis (PCA) [22] was employed to reduce the dimensionality of each of these distributions to two dimensions. The resulting visualization in Figure 6 resulted in 10 data points per subject (annotated as H1, H3, and H5, respectively) corresponding to each of the task categories (represented by specific colors). To calculate the variability in the grasp and manipulation strategies for a given task, a polygon is formed using the projection of each subject for the corresponding task as the vertices. For a given $N$ number of subjects, the inter-subject variability $\mu$

TABLE III

INTER-SUBJECT VARIABILITY OF THE VARIOUS GRASP CLASSIFICATION, GRASP TYPES, AND MANIPULATION TYPES.

| Task Categories | Kitchen Tasks | Inter-subject variablility of | | |
|---|---|---|---|---|
| | | Grasp classification | Grasp type | Manipulation type |
| Cleaning | Cleaning Counter | 0.009177 | 0.002502 | 0.000226 |
| | Trash | 0.008876 | 0.002094 | 0.010818 |
| Cooking | Appliance | 0.000033 | 0.000328 | 0.005291 |
| | Breakfast | 0.034791 | 0.006853 | 0.002828 |
| | Dinner | 0.009028 | 0.000165 | 0.000082 |
| | Lunch | 0.023629 | 0.000294 | 0.003236 |
| Dishwasher | Loading Dishwasher | 0.011449 | 0.002739 | 0.000506 |
| | Unloading Dishwasher | 0.006309 | 0.003409 | 0.011469 |
| Stocking | Fridge | 0.024929 | 0.033466 | 0.002103 |
| | Pantry | 0.052560 | 0.022925 | 0.000308 |

is the area of the polygon given by Eq. 1.

$$\mu = \frac{1}{2}\sum_{k=1}^{N} |(x_k y_{k+1} - x_{k+1} y_k)| \quad (1)$$

where $x$ and $y$ are the 2D coordinates in the dimensionality reduced space. The area of the polygon provides the similarity or variation across the subjects. The variance results of the task categories are detailed in Table III.

The variability is higher if the area of the polygon is higher, indicating that the subjects employed different strategies from each other for the completion of the specific task. It can be noted from the table that a higher variation in the grasp types usually is complemented by a high variance of grasp classification. This could mean that the subjects employ different grasp types classified under different categories (power, precision, intermediate) for successful completion. While the first subject might employ a rigid power grasp, the following subject might use a precision grasp to restrict the motion of a given object. This difference in grasp strategies might stem from a personal preference as well as environmental conditions such as the location of the object. Irrespective of the cause, these tasks require attention during the further data collection stages to determine the ideal strategies to be transferred to robot grippers and hands. Another common observation is that the grasping and manipulation strategies vary inversely with each other in most cases. For example, the stocking of the pantry and fridge had a high variability of grasp types across the subjects. This can be attributed to the different objects/groceries being stocked in each kitchen requiring a different grasping strategy. However, these tasks only involved picking up the groceries and placing them on the pantry/fridge shelf. Hence the manipulation strategy used by all the subjects is identical, resulting in a low variance. On the other hand, all subjects used similar grasping strategies for unloading a dishwasher, but the manipulation strategies varied widely. Similar variation can also be observed in the clearing of trash where the object to be grasped remained identical with different manipulation strategies being employed.

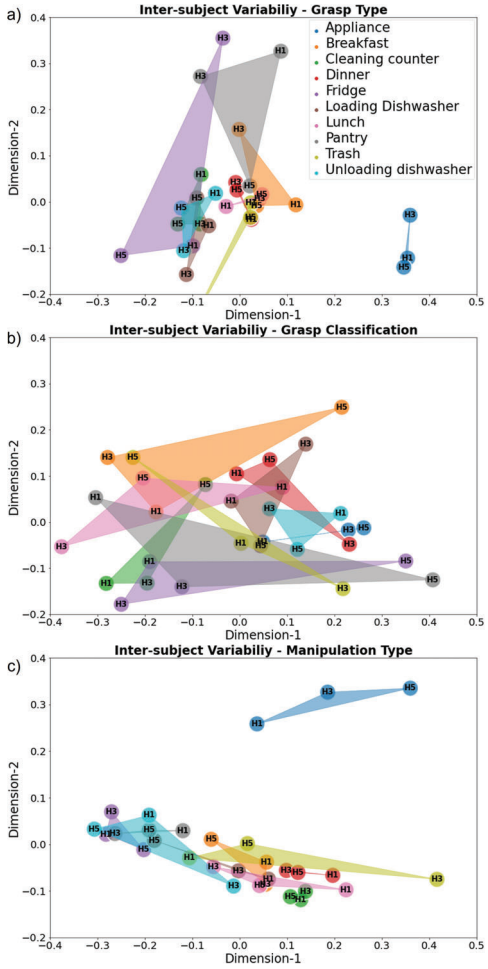


Fig. 6. The inter-subject variability in the a) grasp types, b) grasp classification, and c) manipulation strategies employed for the completion of the various kitchen tasks are calculated from the 2d dimensionality reduced projection of each subject.

The primary focus of this data collection and analysis is to identify the key skills employed by humans for the successful execution of various task categories in a kitchen environment. This would enable the identification and transfer of key skills for the development of robotic grippers and hands that can perform on par with humans as well as collaborate with humans. The tasks with a low variance indicate a standard strategy being employed by all subjects and are directly transferable to robots. Tasks such as cleaning the counters and loading a dishwasher have a very small variance across the subjects, and the grasping/manipulation strategies employed for these tasks can be used by the robotic end-effectors. On the other hand, the tasks with a high inter-subject variance, such as cooking require data collection from a number of subjects to determine the ideal strategy. And determining the cause for the variation across subjects can also provide insights for the selection of optimal grasping and manipulation strategies by considering other parameters such as the speed of execution and accuracy. For example, if each subject uses different strategies for the same objects in similar conditions, the strategy that enables the fastest task execution and higher accuracy can be chosen over others.

Figure 5 presents the inter-subject variability for grasp type and manipulation type for a given task color-coded based on average task completion time. Subject 3 employed "Large diameter grasp" for a majority of the fridge loading task as opposed to subject 1 and subject 5, who performed most of the tasks using "precision sphere" grasp type. It can be seen from Figure 5a that subject 3 had the fastest execution time indicating the effectiveness of "Large diameter" grasp type in loading the fridge. Similarly, in Figure 5b, subject 1 completed unloading the dishwasher faster than other subjects by combining manipulation of grasped objects (with fixed contacts), non-prehensile manipulation of objects (pushing or moving the objects without grasping), and grasp conversion manipulation. While subject 2 and subject 3 were slower as they did not employ non-prehensile manipulation of the objects resulting in a slower completion time. Hence, it is beneficial to impart non-prehensile manipulation strategies to robots to achieve a faster task execution.

Thus, the analysis helps select the optimal strategies employed by the various subjects. It also provides directions for the specific data to be collected and key attributes that require focus during data collection to improve the skill transfer and collaboration between humans and robots.

## V. Conclusion

In this study, we identified key attributes necessary for efficient skill transfer of kitchen tasks execution in a dynamic and unstructured home environment. A comprehensive dataset containing more than 10,000 kitchen activities annotated with 24 attributes each was created from videos of multiple users performing the tasks in kitchen environments. The dataset also includes over 7 hours of high-definition egocentric and supplementary videos of task execution. The study also proposes machine learning techniques capable of extracting the grasp type, the object being manipulated, and the hand performing the grasp from the videos. The efficiency of the proposed framework to partially automate the annotation process by improving the annotation time and saving human resources has been demonstrated.

We further employed a K-means clustering algorithm to identify patterns of critical grasping and manipulation strategies associated with specific kitchen task groups such as stocking, cooking, dishwashing, and cleaning. The clusters enabled us to visualize that a significant part of the kitchen activities can be executed using a limited number of grasps without the need for complex in-hand manipulation. Inter-subject variability was calculated to determine the commonly employed attributes across various subjects that can be transferred to the robots. It also highlights the tasks with high variability between subjects that require attention during further data collection stages. The results presented in this paper can guide and provide key information toward a more efficient and meaningful data collection process that will enable new robotic applications.

## REFERENCES


[1] E. Martinez-Martin and A. P. del Pobil, "Personal robot assistants for elderly care: an overview," *Personal assistants: Emerging computational technologies*, pp. 77–91, 2018.

[2] R. Alqasemi, S. Mahler, and R. Dubey, "Design and construction of a robotic gripper for activities of daily living for people with disabilities," in *2007 IEEE 10th International Conference on Rehabilitation Robotics*. IEEE, 2007, pp. 432–437.

[3] K. Junge, J. Hughes, T. G. Thuruthel, and F. Iida, "Improving robotic cooking using batch bayesian optimization," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 760–765, 2020.

[4] T. X. N. Pham, K. Hayashi, C. Becker-Asano, S. Lacher, and I. Mizuuchi, "Evaluating the usability and users' acceptance of a kitchen assistant robot in household environment," in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2017, pp. 987–992.

[5] Moley, "Moley robotics > the world's first robotic kitchen." [Online]. Available: https://moley.com/

[6] E. Cha, J. Forlizzi, and S. S. Srinivasa, "Robots in the home: Qualitative and quantitative insights into kitchen organization," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, 2015, pp. 319–326.

[7] C. Mason, K. Gadzicki, M. Meier, F. Ahrens, T. Kluss, J. Maldonado, F. Putze, T. Fehr, C. Zetzsche, M. Herrmann *et al.*, "From human to robot everyday activity," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 8997–9004.

[8] I. M. Bullock, J. Z. Zheng, S. De La Rosa, C. Guertler, and A. M. Dollar, "Grasp frequency and usage in daily household and machine shop tasks," *IEEE transactions on haptics*, vol. 6, pp. 296–308, 2013.

[9] C. R. G. Dreher, M. Wächter, and T. Asfour, "Learning object-action relations from bimanual human demonstration using graph networks," *IEEE Robotics and Automation Letters*, vol. 5, no. 1, pp. 187–194, 2020.

[10] B. Jia, Y. Chen, S. Huang, Y. Zhu, and S.-c. Zhu, "Lemma: A multi-view dataset for learning multi-agent multi-task activities," 2020. [Online]. Available: https://arxiv.org/abs/2007.15781

[11] D. Damen, H. Doughty, G. M. Farinella, A. Furnari, E. Kazakos, J. Ma, D. Moltisanti, J. Munro, T. Perrett, W. Price *et al.*, "Rescaling egocentric vision: collection, pipeline and challenges for epic-kitchens-100," *International Journal of Computer Vision*, vol. 130, no. 1, pp. 33–55, 2022.

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

[13] 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.

[14] A. Fathi, Y. Li, and J. M. Rehg, "Learning to recognize daily actions using gaze," in *European Conference on Computer Vision*. Springer, 2012, pp. 314–327.

[15] E. Nicora, G. Goyal, N. Noceti, A. Vignolo, A. Sciutti, and F. Odone, "The moca dataset, kinematic and multi-view visual streams of fine-grained cooking actions," *Scientific Data*, vol. 7, no. 1, pp. 1–15, 2020.

[16] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," 2017. [Online]. Available: https://arxiv.org/abs/1703.06870

[17] I. M. Bullock, R. R. Ma, and A. M. Dollar, "A hand-centric classification of human and robot dexterous manipulation," *IEEE transactions on Haptics*, vol. 6, no. 2, pp. 129–144, 2012.

[18] T. Feix, J. Romero, H.-B. Schmiedmayer, A. M. Dollar, and D. Kragic, "The grasp taxonomy of human grasp types," *IEEE Transactions on human-machine systems*, vol. 46, no. 1, pp. 66–77, 2015.

[19] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C. Chang, M. G. Yong, J. Lee, W. Chang, W. Hua, M. Georg, and M. Grundmann, "Mediapipe: A framework for building perception pipelines," *CoRR*, vol. abs/1906.08172, 2019. [Online]. Available: http://arxiv.org/abs/1906.08172

[20] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022. [Online]. Available: https://arxiv.org/abs/2207.02696

[21] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft coco: Common objects in context," 2014. [Online]. Available: https://arxiv.org/abs/1405.0312

[22] G. Ivosev, L. Burton, and R. Bonner, "Dimensionality reduction and visualization in principal component analysis," *Analytical chemistry*, vol. 80, no. 13, pp. 4933–4944, 2008.