Understanding Human Manipulation with the Environment: A Novel Taxonomy for Video Labelling

Understanding human manipulation with the environment: a novel taxonomy for video labelling

Visar Arapi[1], Cosimo Della Santina[2,3], Giuseppe Averta[4,5,6], Antonio Bicchi[4,5,6], and Matteo Bianchi[5,6]

***Abstract*—In recent years, the spread of data-driven approaches for robotic grasp synthesis has come with the increasing need for reliable datasets, which can be built e.g. through video labelling. To this goal, it is important to define suitable rules to characterize the main human grasp types, for easily identifying them in video streams. In this work, we present a novel taxonomy that builds upon the related state of the art, but it is specifically thought for video labelling. It focuses on the interaction of the hand with the environment and accounts for pre-contact phases, bi-manual grasps as well as non-prehensile strategies. This study is complemented with a dataset of labelled videos of subjects performing activities of daily living, for a total of nine hours, and the description of `MatLab` tools for labelling new videos. Both hands were labelled at any time. We used these labelled data for performing a preliminary statistical description of the occurrences of the here proposed class types.**

***Index Terms*—Datasets for Human Motion, Bimanual Manipulation, Deep Learning in Grasping and Manipulation, Dexterous Manipulation**

I. Introduction

The human example has always been the golden standard and a rich source of inspiration for robotic grasping and manipulation. Not surprisingly, a significant transdisciplinary effort has been devoted to analyze those strategies that humans most often use during the interaction with objects, and classify them in *taxonomies* [1]. The first notable classification of the human-hand versatility was proposed by Schlesinger in [2]. This taxonomy focused on object-contact patterns. In contrast, Napier [3] suggested that a grasp pattern must be identified by considering the intention of the action, and introduced the well-known distinction between *power grasp* and *precision grasp*. Kamakura et al. [4] built and expanded upon Napier's classification to a total of 14 hand patterns.

Cutkosky proposed a more structured taxonomy [5], by leveraging on the concept of *virtual finger* [6] - see Appendix for more details - as well as on the combination of analytic measures (grasp isotropy, force, compliance) to describe and generalize the grasp types. Cutkosky's taxonomy has a hierarchical tree-like structure, where power grasps are classified into nine types and precision grasps into seven types. Following the example of these seminal papers, many others scholars have looked into further extending the range of strategies considered, to improve the granularity of the description [7], [8]. Feix et al. [9] reviewed all the precedent descriptive efforts, and combined them to introduce a single coherent classification: the *GRASP* taxonomy. The organization of this classification has a table-like structure - where the rows are devoted to describe the position of the thumb, while the columns enable to distinguish between power, intermediate and precision grasps; the opposition type; and the virtual finger configuration. Recently, in [10] the authors defined the taxonomy classes moving from the observation of the kinematic measurements and the muscular activation patterns. In [11] the authors proposed a method to augment taxonomies for everyday grasps in action, introducing non prehensile and bimanual tasks such as twisting.

These taxonomies have found fertile applications in different domains [12], which range from the experimental validation of sensing devices [13], to the inspiration for the design and control of new robotic hands [14] and artificial grasping and manipulation strategies [15], just to cite a few. An important application of taxonomies is to provide a support for video labelling [16]. This kind of activity has become of fundamental importance with the spread of machine and deep learning tools for robotic manipulation purposes. To this aim it is fundamental to provide high-quality datasets for model learning and evaluation and to devise data-driven approaches for grasp synthesis. Under this regard, RGB videos (image sequences) represent the principal type of datasets on object manipulation available in literature [17]. In [18] the authors discussed how to predict the grasp type that a human would use to grasp an object from a single RGB image. In [19] a similar strategy was used to implement human-like grasps using anthropomorphic robots. In [20] a framework for learning grasp-manipulation-release tasks from videos of humans performing the task was presented.

However, the taxonomies proposed so far were not developed with the explicit goal of video labelling and hence they have shortcomings when used for this purpose. For example, they do not have the right trade-off between granularity and ease of implementation and usually discard some important aspects that are present in hand-centered video material, such as bimanual grasps. For these reasons, we propose a new taxonomy that was specifically developed for enabling the labelling of human grasping videos. This taxonomy can account for pre-grasp phases, bimanual grasps [21], nonprehensile manipulation [22], and environmental exploitation events [23]. As a first application of the proposed classification, this paper also provides a labelled dataset and the description of a logic workflow complemented with `MatLab` tools for labelling new videos, together with a preliminary statistical description of the occurrences of the proposed classes. To quantify how much the labelling outcomes obtained following our taxonomy depends

This work has been partially supported by the EU H2020 Research and Innovation program under grant agreement no. 732737 (Iliad), and no. 871237 (Sophia) and by the Italian Ministry of Education and Research (MIUR) in the framework of the CrossLab project (Departments of Excellence). [1] Department of Smart Systems Technologies in the Control of Networked Systems Group, Alpen-Adria-Universitat Klagenfurt, 9020 Klagenfurt, Austria. [2] Cognitive Robotics Department, Delft University of Technology, 2628 CD Delft, The Netherlands. [3] Institute of Robotics and Mechatronics, German Aerospace Center (DLR), Oberpfaffenhofen, Germany. [4] Soft Robotics for Human Cooperation and Rehabilitation, Fondazione Istituto Italiano di Tecnologia, via Morego, 30, 16163 Genova, Italy. [5] Centro di Ricerca "Enrico Piaggio", Università di Pisa, Largo Lucio Lazzarino 1, 56126 Pisa, Italy. [6] Dipartimento di Ingegneria dell'Informazione, Università di Pisa, via G. Caruso 16, 56122 Pisa, Italy

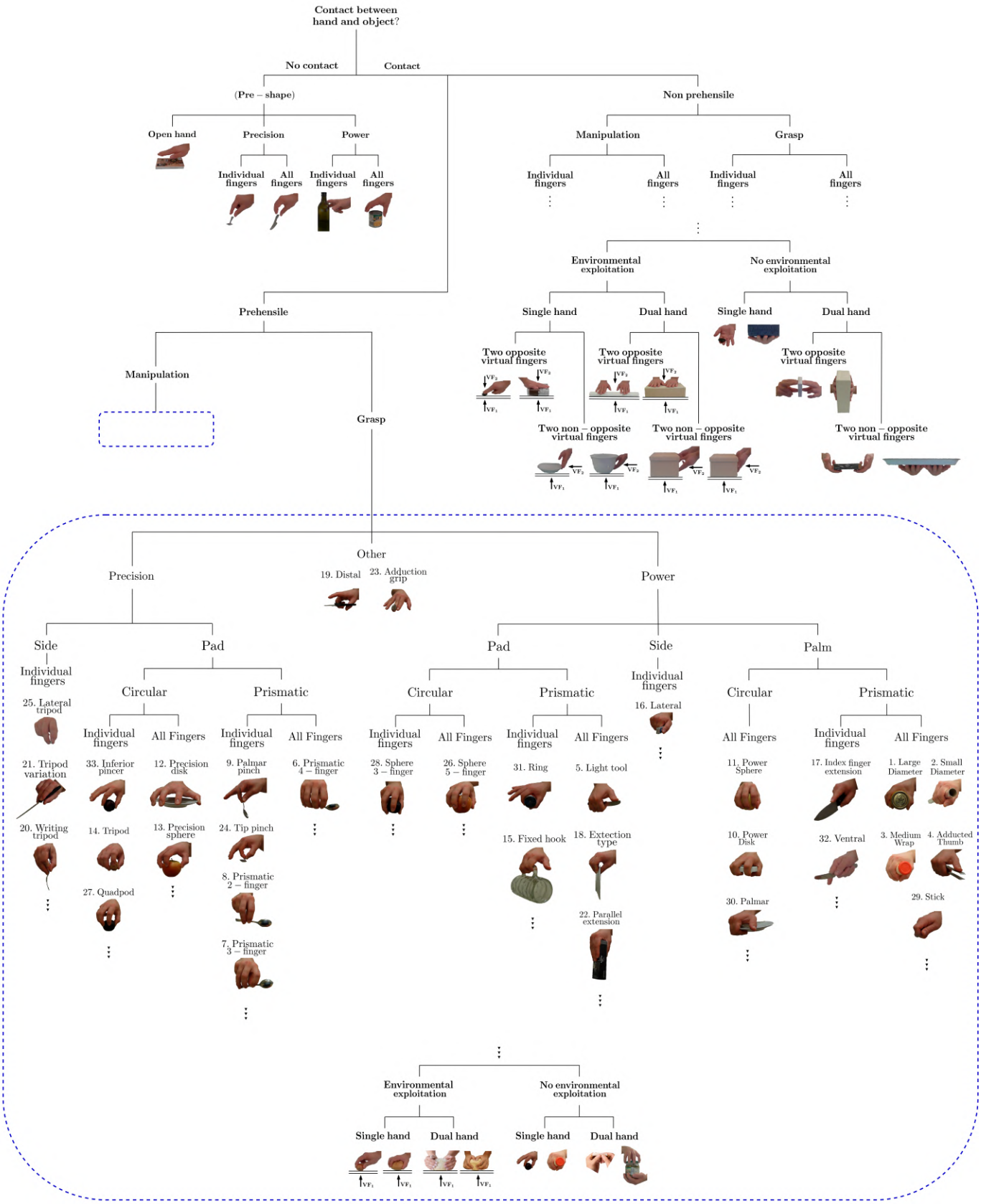

Figure 1. The proposed comprehensive taxonomy. Hand actions are classified following the procedures discussed in the main text. We associate photos to each leaf of the tree to enable a visual interpretation of the taxonomy outcomes. More specifically, in the *Prehensile* and *Non prehensile* cases the leaves contain two examples to represent both the action performed according to the *Individual fingers* and the *All fingers* descriptors Note that some parts of the tree are repeated across branches. Therefore we decided to show them just once for the sake of readability. We showed only the prehensile grasp sub-categorization and recalled it with a dashed blue rectangle in the prehensile manipulation sub-categorization since it is the same in both the cases.

on the specific labeller, we asked two users to label the same amount of data and statistically compared the results using Cohen's kappa score [24]. With a score of 0.82 computed over 177 grasping/manipulation actions extracted from cooking videos, we can suggest a reduced effect of users' subjectivity on the application of our taxonomy. Of note, this point will deserve further investigation and analysis in future work.

II. Overall Taxonomy Description

The taxonomy is organized as a hierarchical tree, as depicted in Fig. 1. The taxonomy is implemented considering one single hand at a time, the *observed* hand. If two hands appear simultaneously in a video, then an independent characterization is applied to each of them. Of note, if the two hands synergistically cooperate in the same goal-directed grasp action, each characterization is complemented by additional descriptors that identify between-hand cooperation. The first distinction in our taxonomy is between *Contact* vs *No Contact*. To the latter is associated the macro-category related to *Pre-shape*, while to the former the choice is between the two macro-classes *Prehensile* vs. *Non-prehensile*. Each of these macro classes is then broken up in smaller classes that are detailed in the following sections. The *Pre-shape* comprises the sub-classes open hand, precision and power, the latter two are further divided according to the all fingers vs. individual fingers distinction. The *Non-prehensile* type is divided in individual vs all fingers, and again all the branches are subdivided w.r.t. the presence and nature of the environmental exploitation and the number of hands involved (i.e. single vs dual hand, the latter refers to between-hand cooperation). The *Prehensile* type is subdivided by following a grasp-manipulation taxonomy that is described in the dedicated Fig. 2. Then the presence of environmental exploitation and the number of hands involved is considered. The total number of leaves is 124, but the descriptors are only 27 -as listed in Tab. I with the naming used in the video labeller GUI- and can be combined to generate all the classes. The Appendix of this paper provides a short description of the main terms used in the taxonomy description.

III. Pre-shape

During the reaching phase the hand already starts preparing for the task, by assuming the proper configuration for the strategy that will be implemented. We introduce five pre-shape types in our taxonomy, each of them starts when there is an intention to grasp, and ends when the hand has contacted with or enclosed the object. These types are shown in the top left part of Fig. 1.

1) Open Hand: the hand is approaching the object without changing its shape. Generally, this pre-shape leads to a *non-prehensile* action and the hand will act as a single virtual finger when touching the object. This type is identified by the descriptor *PRE_Open_Hand* reported in Table I.

2) Precision - individual fingers: only few tips of the distal phalanxes are moved forward compared to the other parts of the hand. This pre-shape then results in a *prehensile* precision grasp or manipulation action. This type is identified by the descriptor *PRE_Precision_Partial* in Table I.

3) Precision - all fingers: the same as for *Precision - individual fingers*, but all the fingers are moved so to prepare a contact with the object. This type is identified by the descriptor *PRE_Precision_Whole* in Table I.

4) Power - individual fingers: the hand is shaped so to expose an extended contact area to object to be grasped, which usually comprises all the inner parts of the fingers and the palm. Only few fingers are moved forward with respect to the rest of the hand, which will then establish a contact with the object. This pre-shape action often results into a *prehensile* power grasp or manipulation action. This type is identified by the descriptor *PRE_Power_Partial* in Table I.

5) Power - all fingers: the same as for *Power - individual fingers*, but all fingers are simultaneously approaching the object. This type is identified by the descriptor *PRE_Power_Whole* in Table I.

IV. Non-prehensile

Non-prehensile manipulation strategies are very common in humans. Nonetheless, relatively little attention has been devoted in the existing taxonomies. A single category for non-prehensile tasks is proposed in [5]. A more extensive non-prehensile classification is provided in [25]. The authors divided the non-prehensile task in two main classes: *motion* and *non-motion*, respectively.

In developing our non-prehensile categorization (see Fig. 1), we first distinguish between *Manipulation* and *Grasp*, depending whether the finger configurations do change with respect to the object or not. Then, we identify two further categories for each of the two classes: *Individual Fingers* and *All Fingers*. This last distinction carries the information of whether the hand is partially or fully involved in the execution of the task. Finally, for each of the four classes (which correspond to the four descriptors *NON_PREHENS_** of Table I, where *PARTIAL* and *WHOLE* apply to *Individual Fingers* and *All Fingers*, respectively), we propose a comprehensive sub-categorization that aims at characterizing the exploitation of the environment, and/or the presence and the nature of the bi-manual operation aka between-hand cooperation. This is shown in the bottom left of Fig. 1.

1) Environment exploitation: in this type of classes, which is identified by the descriptor *ENVIRONMENT_ EXPLOITATION* in Table I, the environment has a functional role in the grasping or manipulation strategy. In these cases, the environment usually acts as a virtual finger (VF1) in addition to the one provided by the hand (VF2). We can distinguish two scenarios: *i*) the force generated by the environment VF1 is in opposition to VF2 (*Two opposite virtual fingers* in Fig. 1); *ii*) the two virtual fingers VF1 and VF2 are not in opposition (*Two non-opposite virtual fingers* in Fig. 1).

2) Bi-manual operation: this category is also referred to as Dual Hand see Fig. 1 and implies a cooperation between the two hands. In Table I the corresponding descriptor is *HANDS_COOPERATION*. In this case, the two hands play the role of the two virtual fingers. As for the *Environment exploitation* type, we can further distinguish between *Two opposite virtual fingers* - resulting in a pinch-like grasp, and *Two non opposite virtual fingers*.

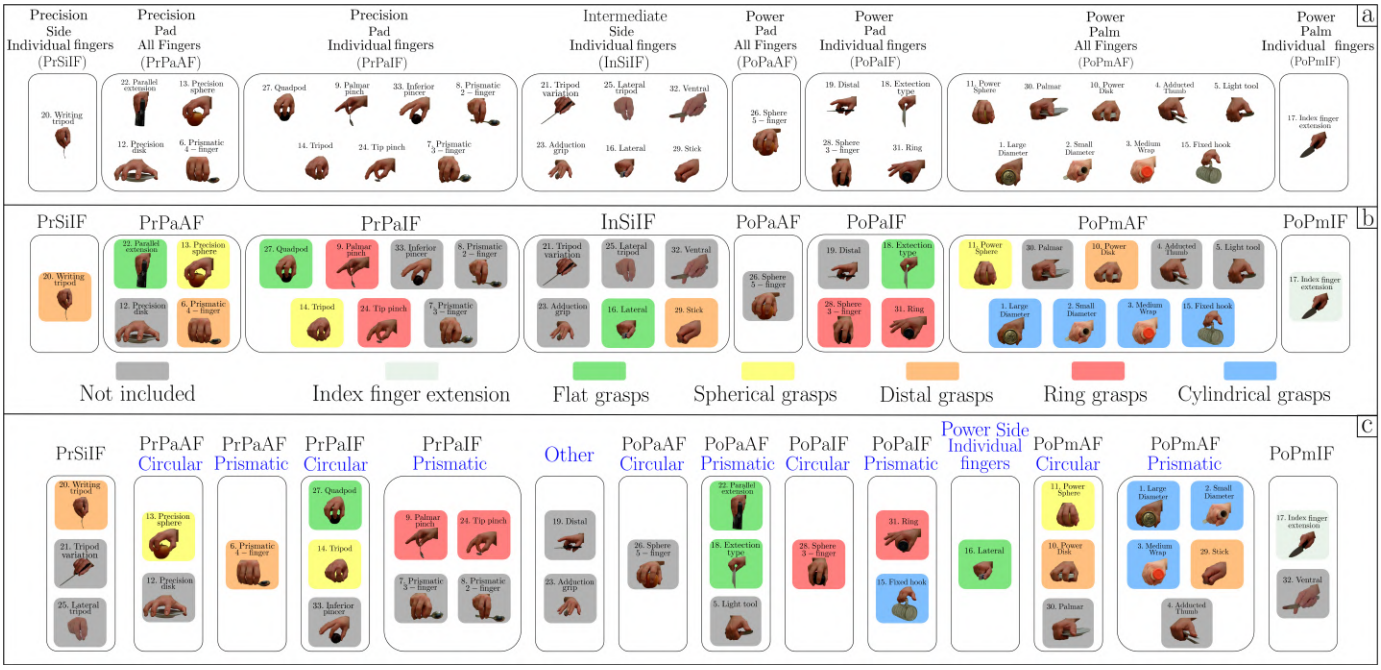

Figure 2. a) We reduced the 33 grasps of the GRASP taxonomy to 8 categories. b) Relationship between the six groups defined in [11] and represented using the color-code light green, green, yellow, orange, red, blue, with the eight macro-classes in sub-figure a); the 13 grasps of the Feix's taxonomy not considered in [11] are highlighted in gray. c) The 14 categories of our taxonomy – and the relation with [10] and [11]: *PrSiIF*; *PrPaAF Circular*; *PrPaAF Prismatic PrPaIF Circular*; *PrPaIF Prismatic*; *Other*; *PoPaAF Circular*; *PoPaAF Prismatic*; *PoPaIF Circular*; *PoPaIF Prismatic*; *Power Side Individual Fingers*; *PoPmAF Circular*; *PoPmAF Primsatic*; *PoPmIF*.

V. Prehensile

As for the *Non-prehensile* case, the first distinction is between *Manipulation* and *Grasp*, which correspond to the descriptors *PREHENSILE_MANIPULATION* and *PREHENSILE_GRASP* in Table I.

To devise the classes of our taxonomy related to the prehensile case, our starting point was the well-known GRASP taxonomy [9]. A main obstacle that prevents in our opinion an extensive application of the GRASP taxonomy to video labelling is the need for a precise recognition of the thumb abduction/adduction position. This is challenging since in video material the thumb is often only partially visible. Another critical point is related to the identification of the exact number of fingers involved during the task execution. To reach a trade-off between the accuracy and ease of implementation, the 33 grasps were reduced to 8 categories by implementing the following actions: i) merging the grasps that are within the same cell in [9] into a unique grasp leaf, thus reducing the number of grasps from 33 to 17; ii) removing the thumb abduction/adduction differentiation (from 17 to 14); iii) introducing two new categories, *Individual Fingers* and *All Fingers* to facilitate the description on the number of fingers acting as virtual fingers; this overrides the previous classification based on the specific fingers involved in the action, reducing the number of 14 grasps to 8. *Individual fingers* and *All fingers* distinction is equivalent to the one already discussed in Sec. IV. Fig. 2-a depicts how the original 33 grasps are included into our reduced 8 categories. Please refer to 2-a for specific names of each category, and to the Appendix for their definition (re-adapted from [9]).

In [10] the authors considered 20 of the 33 grasps proposed in [9] and categorized them in: *1) flat grasps* characterized by an elongated positioning of the palm; *2) distal grasps* characterized by the strong involvement of distal phalanxes; *3) cylindrical grasps* strongly linked to the shape of cylindrical objects; *4) spherical grasps* strongly linked to the shape of spherical objects; *5) ring grasps* characterized by the involvement of the thumb and index finger only; *6) index finger extension*. This characterization moved from the quantitative analysis of kinematic and electromyographic data. We evaluated the relationship between the six groups, which we represented using the color-code light green, green, yellow, orange, red, blue (see Fig.2-b), defined in [10] with the eight macro-classes we identified from Feix's taxonomy. This is shown in Fig.2-b. The 13 grasps of the Feix's taxonomy that are not considered in [10] are highlighted in gray. It is worth noticing that the six classes in [10] do not cluster and differentiate each other according to our eight categories. For this reason, we decided to modify them implementing the following actions i) we introduced the *Other* sub-category (to group grasps that cannot be easily distinguished each other i.e. #19, #23) and the *Power Side Individual Fingers* category corresponding to the lateral grasp #16; ii) we defined the *prismatic* and *circular* sub-categories (see Appendix) to characterize the shape of the object for the *PoPmAF*, *PoPaIF*, *PoPaAF*, *PrPaIF* and *PrPaAF* categories; iii) we removed the *Intermediate* category, which is typically not considered in state-of-the-art literature as discussed in [9]. The final configuration of our prehensile taxonomy consists of 14 categories which are here omitted for sake of space and are reported in Fig 2-c (with relative descriptors listed in Tab I), together with a visual representation of this sub-taxonomy and how the grasps in [9] and the categorization in [10] are organized w.r.t our 14 categories, which apply to both the grasp and manipulation branch in Fig. 1. Each leaf of the grasp/manipulation taxonomy defined as above is complemented by two additional labels, which are depicted in

Table I
LIST OF THE 27 DESCRIPTORS USED IN VIDEO LABELER GUI (PARTIAL AND WHOLE REFER TO INDIVIDUAL AND ALL FINGERS)

PRE_Open_Hand	PREHENSILE_GRASP	Power_Pad_Partial_Circular
PRE_Power_Partial	PREHENSILE_MANIPULATION	Power_Pad_Whole_Prismatic
PRE_Power_Whole	ENVIRONMENT_EXPLOITATION	Power_Pad_Whole_Circular
PRE_Precision_Partial	HANDS_COOPERATION	Other
PRE_Precision_Whole	Power_Palm_Partial	Precision_Pad_Partial_Prismatic
NON_PREHENS_GRASP_Partial	Power_Palm_Whole_Prismatic	Precision_Pad_Partial_Circular
NON_PREHENS_GRASP_Whole	Power_Palm_Whole_Circular	Precision_Pad_Whole_Prismatic
NON_PREHENS_MANIP_Partial	Power_Side_Partial	Precision_Pad_Whole_Circular
NON_PREHEN_MANIP_Whole	Power_Pad_Partial_Prismatic	Precision_Side_Partial

the bottom of Fig. 1 and correspond to the descriptors *ENVIRONMENT_EXPLOITATION* and *HANDS_COOPERATION* in Table I, as we defined also for the non prehensile case: 1) *Environment exploitation*: although the object is stabilized by the hand, the exploitation of the environment can be required in some specific tasks. For instance, in bread cutting tasks, we first hold the bread firmly, then we press it against the cutting board which provides an additional virtual finger. 2) *Bi-manual operation* or *Dual Hand*: this case includes two scenarios *i*) both hands result in a prehensile mode, e.g. getting a jar lid unstuck; *ii*) the observed hand results in a prehensile fashion while the second hand results in a non-prehensile mode. It is worth noting that since the stability of the object is provided by the observed hand, the resulting movement preserves the prehensibility regardless of the two mentioned scenarios.

VI. PROPOSED DATASET AND LABELLING TOOLS

We report here a dataset of videos where human hands have a prominent role, which we labelled using the proposed taxonomy. Overall it contains approximately 9 hours (531 minutes), and is provided as supplementary material to this manuscript. It is worth stressing that this paper contributes with the labelling only, since the video material was already publicly available online. We consider three sets of videos and employed our taxonomy to annotate them. We generated the first set by collecting cooking videos from YouTube. We call it the Cooking dataset. The second is the Agreement dataset, a subset of the Cooking dataset used as a test bed to evaluate the labeller specific effect on the annotation. The third is the 20BN-something-something dataset [1] V1 [26], which we re-labelled relying on the here proposed taxonomy.

1) Cooking Dataset: Cooking is an activity that elicits a variety of complex hand goal-directed motions. Also, there is an abundance of videos available online of people teaching how to cook recipes, where the hands are often clearly visible. For our dataset collection, we considered the Youtube playlist of two well-known cooks, then, before selecting the videos we visually inspected them. We collected the videos by considering: *i*) unconstrained third-person view videos of single subjects performing food preparation activities; *ii*) different dish preparation in order to get a variation in hand movements. Within these dishes each cook used different ingredients (bread, beef, tomatoes, cheese, etc.) as well as potential tools (knife, grater, spoon, etc.), resulting in very dissimilar videos. Such features make these videos a good test bed for the application of our taxonomy. Dish preparation

[1]https://20bn.com/datasets/something-something/v1

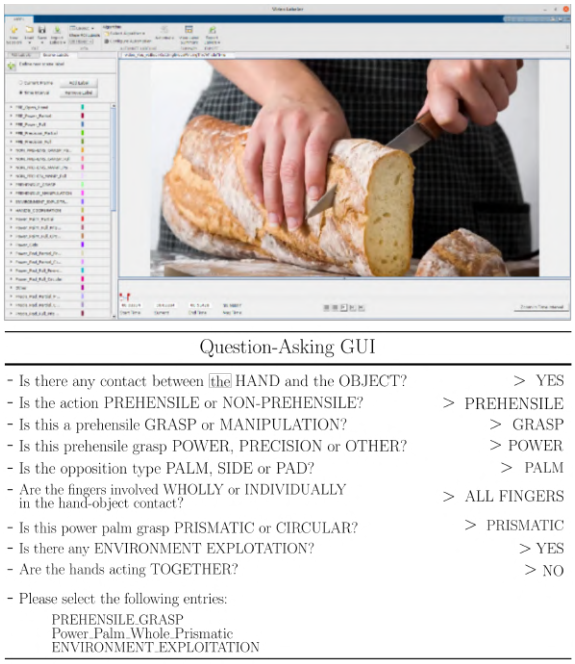

Figure 3. On the top, the annotation GUI we developed in this work. It is based on `MatLab`'s VideoLabeler tool, which we modified to be used with the proposed descriptors listed in Tab. I. On the bottom, a session of the question-asking GUIDE implemented in `MatLab`. In this session example, the GUI interacts with the user identifying the descriptors relative to the right hand of the subject in the top image.

video time varies from 6 to 12 minutes. In total the Cooking dataset includes 25 videos with a total length of more than 3 hours (183 minutes). All videos have a resolution of 1280×720 pixels and a frame rate equal to 24.

2) Agreement Dataset: This dataset was established to evaluate the effect of labeller's subjectivity on the annotation analysis based on our taxonomy. For this purpose, we segmented randomly 177 video intervals among the Cooking dataset videos. The selection was performed manually in order to cover the different instances of the taxonomy and ensure a single hand configuration in each segmented video. The duration of these videos ranges from 1 to 9 seconds.

3) 20BN-something-something (20BN) Dataset: 20BN dataset is a large collection of densely-labeled video clips that show objects and actions performed on them, by humans standing in front of a camera [26]. The videos are about simple, and mostly everyday actions and events. Labels are in textual form and represent detailed information about the objects and actions as well as other relevant information. The dataset is very large and various, and for our purpose we selected out 20 of the 174 labels on the validation dataset.

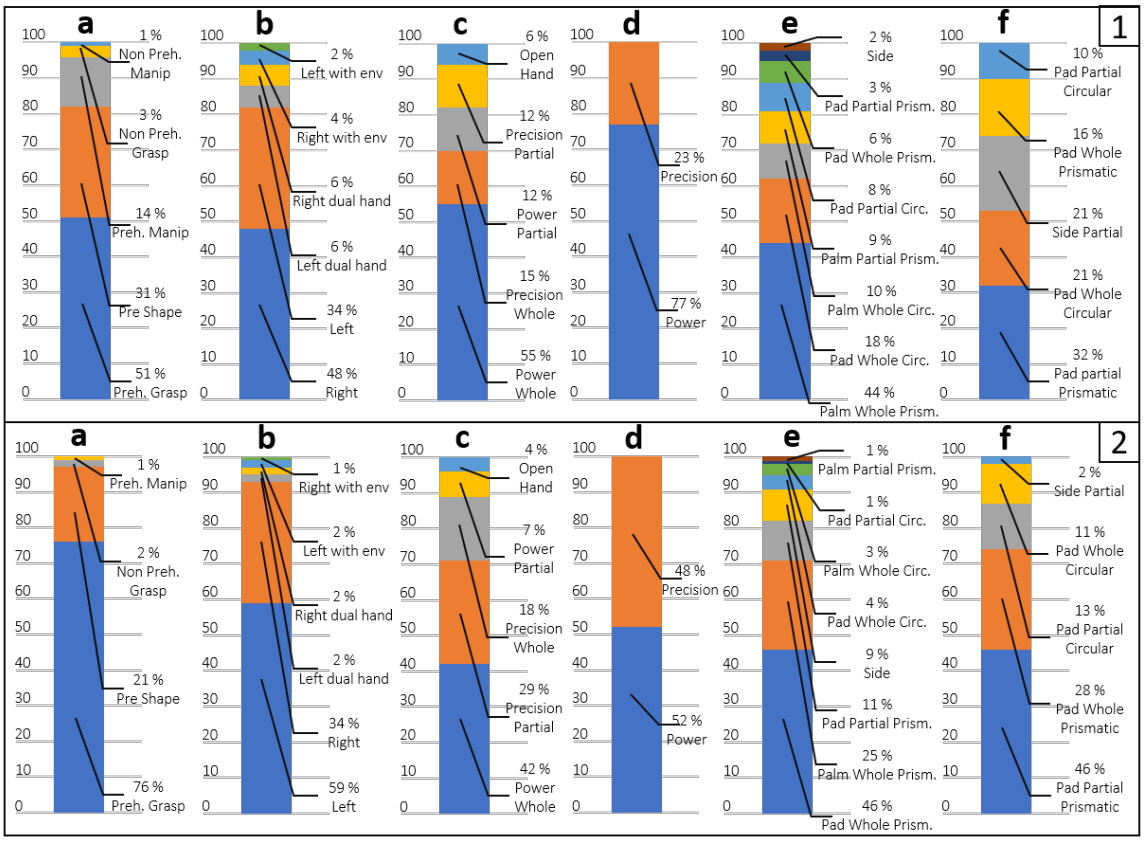

Figure 4. Instances distribution for the Cooking Dataset (1), and, instances distribution for the 20BN Dataset (2). For both the datasets, the following distributions are derived: a) Distribution of the total instances; b) Right and left hand instances including environmental exploitation and dual hand executions; c) Pre-shape instances grouped by using the 5 pre-shape types defined in our taxonomy; d) Distribution of power and precision grasp instances; e) Power grasp types; f) Precision grasp types.

For the selection, we visually inspected some of the videos belonging to each of the 174 action categories. The aim of the inspections was to avoid action categories where no hand-object contacts occur, as well as to select those categories that elicits a variety of hand configurations (i.e. holding, closing, moving, etc.). We considered a total of 4943 videos (with duration ranging from 2 to 6 seconds) with a total length of approximately 6 hours (348 minutes). The videos have different resolutions and a frame rate equal to 12.

A. Data Annotation and Labelling tool

Two experienced researchers (Labeler1 and Labeler2) with an engineering background and expert in the human grasping and manipulation literature annotated the Agreement dataset. While for the Cooking and 20BN datasets only Labeler1 was employed for the annotations. The labelers manually annotated the RGB videos, using the *VideoLabeler* tool of `MatLab`. We customized the GUI to our purpose by including label descriptors previously defined in Tab. I, combination of which identifies a specific leaf of the taxonomy. This modified GUI is available for download as supplementary material to this manuscript. During the labelling session, the labeler can start and stop the video manually, as well as slide the video frame-by-frame. The start and stop timestamps of the corresponding video segment are annotated, and, when the whole video is analyzed we save both the GUI session and the label structure which has the same frame rate and the same length as the analyzed video has. In the case of *preshaping*, the start frame of the segmentation is identified as the intention of the subject to reach an object, while the stop frame is determined when the hand touches the object. In the case of hand object contact, the start frame of the new interval is accordingly defined once the hand contact arises. The stop frame is recorded when one of the two following conditions is verified, *i*) the hand object contact is still maintained, but, the hand movement transits into another category (which also becomes the start frame of the subsequent interval); *ii*) the subject releases the object.Quick hand movement transitions lasting less than 0.2 sec are not considered, since they produce blurry images, with consequent hand shape uncertainty. This GUI was complemented with an additional tool, implemented with the GUIDE tool in `Matlab`, which was used to interact with the annotator by presenting information about the hand movement. This second GUI is based on Fig. 1 and asks questions which help the selection of the proper leaf. Each question provides details on the keywords used - the same defined in the Appendix - to discriminate between the tree paths, and is complemented with pictures of the different cases (the same reported in Fig. 1). The question-asking GUI allows the user to advance in any direction of the branches of the taxonomy (with the possibility to step back) and eventually terminates in one of the leafs of the taxonomy, which is then the input of the VideoLabeler GUI. Both the GUI tools are released together with this manuscript, with the hope that the proposed approach could be expanded by other contributors. Fig. 3 (top) shows the screenshot of the

VideoLabeler GUI, while Fig. 3 (bottom) shows the outcomes of a session example of the second tool GUI.

B. Inter-subject Agreement and Statistical Analysis

1) Inter-subject Agreement: To quantify how much the labelling with our taxonomy may change depending on the labeler, we asked two users to label the same dataset and statistically compared the two outcomes. For this purpose we considered the pairs of 177 annotations of the Agreement dataset. In particular, both the labelers distinguished correctly in the same way: *i*) 4 non-prehensile actions; *ii*) 173 prehensile actions (5 as manipulations and 168 as grasps); *iii*) 4 dual-hand cooperation (2 for the non prehensile case and 2 for the prehensile one); *iv*) 7 environmental exploitation (2 for the non prehensile case and 5 for the prehensile one). Furthermore, the sub-classification was the same for the non-prehensile actions and prehensile manipulations, except for the prehensile grasps. In the last case, we noticed that the main source of errors for both the labelers is related to the area of contact between the fingers/palm and the object. Instead, other characteristics such as number of fingers during the contact and the shape of the object are both correctly identified in most of the cases. To provide a measure of agreement of the labelers among the 14 prehensile grasp categories, we used the Cohen's kappa score [27] on the annotated instances, obtaining $\kappa = 0.82$. Considering that $\kappa = 1$ is the optimal case, we can emphasize the fact that the designed taxonomy seems to ensure a reduced human-related effects on the video labelling. In the following, we give a statistical characterization of the labels that the Labeler 1 associated to the datasets.

2) Cooking Dataset: Of all the 3114 hand configurations identified in the Cooking dataset, 51% were classified as Prehensile Grasps, 31% as Pre-shape, 14% as Prehensile Manipulation, (3%) as Non-Prehensile Grasp and (1%) as Non-Prehensile Manipulation (see Fig. 4-1a). The hand side used to execute the action in the Cooking dataset is well balanced, resulting in a Right hand for the 48% of the cases and Left hand for the 34% (see Fig. 4-1b). Other cases observed are Left with Dual Arm and Right with Dual Arm (6% each), Right with the Environment (4% each) and Left with the Environment (2% each). Note that even if some of these classes have small percentages, together they add up to a non negligible amount of video materials. Considering the Pre-shape types (see Fig. 4-1c), in the Cooking dataset we verified that in the 55% of the cases people performed a Power Whole, while a Precision Whole is executed in 15% of the entries, and Precision and Power Partial occur in 12% of the cases each. Finally, Open Hand label is applied in 6% of the cases. For the Prehensile Grasps (see Fig. 4-1d), instead, we observed a Power grasp in 77% of cases, and a Precision grasp in the remaining 23%. Furthermore, Fig. 4-1e and Fig. 4-1f depict how the power and precision grasp leafs are distributed, respectively.

3) 20BN Dataset: Of all the 5168 hand postures identified in the 20BN dataset, 76% were classified as Prehensile Grasps, 21% as Pre-shape configurations, 2% as a Non Prehensile Grasp and only 1% as Prehensile Manipulation (see Fig. 4-2a). More than half of the total instances were executed with the left hand only (59%), while the right hand was used in the 34% of cases. Minor classes are Left and Right with Dual Hand (2% each) and Left and Right with Environment (2% and 1% respectively, see Fig. 4-2b). Regarding the Pre-shape instances (Fig. 4-2c), the 20BN dataset is composed of Power Whole in the 42% of cases, Precision Partial in 29% and Precision Whole in 18%. Minor classes are Power Partial (7%) and Open Hand (4%). Regarding the Prehensile grasp, instead, this almost equally divided between Power (52%) and Precision (48%) grasps (see Fig. 4-2d). Furthermore, Fig. 4-2e and Fig. 4-2f depict how the power and precision grasp leafs are distributed, respectively.

4) Comparison with Literature: A comparison between our distributions and the ones proposed in literature is not immediate, because of the different hand taxonomies. In [16] and [24] the authors adopted the GRASP taxonomy to annotate their datasets -food preparation/cleaning, housekeeping and laundry in [16], housekeeping and mechanist in [24]. In the GRASP taxonomy there are three macro-categories (power, intermediate and precision) while in ours we have only two (power and precision). Since in ours most of the entries of the intermediate category were incorporated into the power one, to make possible the comparison we summed together the power and intermediate distributions for both [16] and [24]. Hence, in [16] the Power and precision are distributed as (45%, 55%) in the food preparation and cleaning, (67%, 33%) in the housekeeping and, (74%, 26%) in the laundry. In [24] the Power and precision are distributed as (67%, 33%) in the housekeeping and, (41%, 59%) in the mechanist. Power and precision in our annotations are distributed as (77%, 23%) in the Cooking dataset and (58%, 42%) in the 20BN dataset. It is worth noticing that the total number of grasp actions annotated in [16] is 3826, in [24] is approximately 4700, while in our dataset we annotated 8282 instances (3114 for the Cooking dataset and 5168 for the 20BN dataset).

VII. Conclusion

This paper proposed a new grasp and manipulation taxonomy which is wide enough to encompass all actions that are typically recognizable in hand-centered videos. The classes are specified to be accurate, but simultaneously general enough for allowing labelling of standard streams of images - where the hand is not necessarily clearly visible from all sides. We provided an example of labeled dataset which can be downloaded, together with `MatLab` tools that we developed for facilitating the labelling. Preliminary statistical characterization of the dataset has also been provided. We also evaluated the effect of users' subjectivity on the application of our taxonomy asking two labelers as in [11] to annotate the same amount of videos. We found a reduced subjectivity effect although this point will deserve further investigation and analysis in future work. Future work will be also devoted to train deep neural networks for automatic labelling of new videos, and for artificial grasp generation in robotics [19], [28].

Appendix: Glossary

Virtual Finger (VF): a virtual finger is a single functional unit acting as a single imaginary finger, comprised of multiple real physical fingers and/or the palm. They must act in unison to apply a force on the object and against the other virtual fingers. In this work we consider a relax version of this concept, to include environment exploitation and bimanual

tasks. In these cases additional virtual fingers are generated by external components. *Grasp*: this is a kind of object-hand interaction such that, once the contact is established, the fingers' configuration with respect to the object does not change, but in principle the object can still move w.r.t. the hand, such as in sliding. *Manipulation*: similar to Grasp, but the fingers' configuration does change. *Non-Prehensile*: the fingers or the palm involved in the interaction act as a single virtual finger, which is then used to hold, push, or lift the object. *Prehensile*: at least two virtual fingers (VFs) are being applied in opposition against the object's surfaces. The contact forces from the hand alone are therefore sufficient to stabilize the object against gravity and other external forces. *Environment exploitation*: the subject purposefully exploits the environment to perform the task (i.e. there is a contact between the object and the environment). *Bi-manual operation*: both hands are involved during the task execution. *Power*: characterized by large areas of contact between the grasped object and the surfaces of the fingers or palm and by little ability to impart motions with the fingers. This means that movements of the object have to be evoked both by the fingers and the arm. *Precision*: the fingers configuration is such that the object is held generally between the tips of two or more fingers - the thumb is usually involved - and the hand is able to induce intrinsic movements on the object without having to move the arm. *Power/Palm*: the forces generated by the fingers define directions approximately perpendicular to the palm plane. *Power/Side*: the forces generated by the fingers define directions approximately transverse to the palm plane, involving the external lateral side of one or more fingers as it is the case of grasping a credit card. *Power/Pad*: the forces generated by the fingers define directions approximately parallel to the palm plane. *Precision/Pad*: the object is held between the distal phalanxes and direction of forces applied to it are generally parallel to the palm plane. *Precision/Side*: the object is held between the lateral (radial-internal) part of the middle or distal phalanx of the index-middle finger and the pulp of the thumb, as it is the case of holding a pen for writing. *Intermediate/Side*: This category is in an intermediate position between Power and Precision. The palm is no longer included as a contact area. The fingers are generally in moderate flexion and the contact areas include the radial aspect of the index or the middle finger. *Prismatic*: it depends on the shape of the object. Objects that require a prismatic grasp are characterized by the dimension of the cross-section that is significantly smaller than the dimension of its perpendicular side (such as for a long cylinder or hexagonal prism). *Circular*: this case groups round objects and the dimension perpendicular to circular cross-section is less or equal to the cross-section itself.

REFERENCES

[1] T. Iberall, "The representation of objects for grasping," in *Proceedings of the Eighth Cognitive Society Conference*. Cognitive Society, 1986, pp. 547–561.

[2] G. Schlesinger, "Der mechanische aufbau der künstlichen glieder," in *Ersatzglieder und Arbeitshilfen*. Springer, 1919, pp. 321–661.

[3] J. R. Napier, "The prehensile movements of the human hand," *The Journal of bone and joint surgery. British volume*, vol. 38, no. 4, pp. 902–913, 1956.

[4] N. Kamakura, M. Matsuo, H. Ishii, F. Mitsuboshi, and Y. Miura, "Patterns of static prehension in normal hands," *American Journal of Occupational Therapy*, vol. 34, no. 7, pp. 437–445, 1980.

[5] M. R. Cutkosky, "On grasp choice, grasp models, and the design of hands for manufacturing tasks." *IEEE Transactions on robotics and automation*, vol. 5, no. 3, pp. 269–279, 1989.

[6] G. Baud-Bovy and J. F. Soechting, "Two virtual fingers in the control of the tripod grasp," *Journal of Neurophysiology*, vol. 86, no. 2, pp. 604–615, 2001.

[7] S. J. Edwards, D. J. Buckland, and J. D. McCoy-Powlen, *Developmental and functional hand grasps*. Slack, 2002.

[8] I. A. Kapandji, *Funktionelle Anatomie der Gelenke: schematisierte und kommentierte Zeichnungen zur menschlichen Biomechanik*. Georg Thieme Verlag, 2006.

[9] 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.

[10] F. Stival, S. Michieletto, M. Cognolato, E. Pagello, H. Müller, and M. Atzori, "A quantitative taxonomy of human hand grasps," *Journal of neuroengineering and rehabilitation*, vol. 16, no. 1, p. 28, 2019.

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

[12] C. Choi, S. Ho Yoon, C.-N. Chen, and K. Ramani, "Robust hand pose estimation during the interaction with an unknown object," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3123–3132.

[13] S. Sundaram, P. Kellnhofer, Y. Li, J.-Y. Zhu, A. Torralba, and W. Matusik, "Learning the signatures of the human grasp using a scalable tactile glove," *Nature*, vol. 569, no. 7758, pp. 698–702, 2019.

[14] H. Stuart, S. Wang, O. Khatib, and M. R. Cutkosky, "The ocean one hands: An adaptive design for robust marine manipulation," *The International Journal of Robotics Research*, vol. 36, no. 2, pp. 150–166, 2017.

[15] R. Ozawa and K. Tahara, "Grasp and dexterous manipulation of multi-fingered robotic hands: a review from a control view point," *Advanced Robotics*, vol. 31, no. 19-20, pp. 1030–1050, 2017.

[16] A. Saudabayev, Z. Rysbek, R. Khassenova, and H. A. Varol, "Human grasping database for activities of daily living with depth, color and kinematic data streams," *Scientific data*, vol. 5, no. 1, pp. 1–13, 2018.

[17] Y. Huang, M. Bianchi, M. Liarokapis, and Y. Sun, "Recent data sets on object manipulation: A survey," *Big data*, vol. 4, no. 4, pp. 197–216, 2016.

[18] E. Corona, A. Pumarola, G. Alenya, F. Moreno-Noguer, and G. Rogez, "Ganhand: Predicting human grasp affordances in multi-object scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5031–5041.

[19] C. Della Santina, V. Arapi, G. Averta, F. Damiani, G. Fiore, A. Settimi, M. G. Catalano, D. Bacciu, A. Bicchi, and M. Bianchi, "Learning from humans how to grasp: a data-driven architecture for autonomous grasping with anthropomorphic soft hands," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1533–1540, 2019.

[20] N. Wake, R. Arakawa, I. Yanokura, T. Kiyokawa, K. Sasabuchi, J. Takamatsu, and K. Ikeuchi, "A learning-from-observation framework: One-shot robot teaching for grasp-manipulation-release household operations," in *2021 IEEE/SICE International Symposium on System Integration (SII)*. IEEE, 2021, pp. 461–466.

[21] R. Chitnis *et al.*, "Efficient bimanual manipulation using learned task schemas," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 1149–1155.

[22] M. R. Dogar and S. S. Srinivasa, "A planning framework for non-prehensile manipulation under clutter and uncertainty," *Autonomous Robots*, vol. 33, no. 3, pp. 217–236, 2012.

[23] S. Puhlmann, F. Heinemann, O. Brock, and M. Maertens, "A compact representation of human single-object grasping," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 1954–1959.

[24] I. M. Bullock *et al.*, "Grasp frequency and usage in daily household and machine shop tasks," *IEEE transactions on haptics*, vol. 6, no. 3, pp. 296–308, 2013.

[25] I. M. Bullock and A. M. Dollar, "Classifying human manipulation behavior," in *2011 IEEE International Conference on Rehabilitation Robotics*. IEEE, 2011, pp. 1–6.

[26] R. Goyal *et al.*, "The" something something" video database for learning and evaluating visual common sense," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5842–5850.

[27] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[28] V. Arapi, C. Della Santina, D. Bacciu, M. Bianchi, and A. Bicchi, "Deep-dynamichand: A deep neural architecture for labeling hand manipulation strategies in video sources exploiting temporal information," *Frontiers in neurorobotics*, vol. 12, p. 86, 2018.